JUMS

Volume 10, Issue 3, September 2025

Editor:
DOMINIK VAN AAKEN

Advisory Editorial Board:
FREDERIK AHLEMANN
JAN-PHILIPP AHRENS
THOMAS BAHLINGER
MARKUS BECKMANN
SULEIKA BORT
ROLF BRÜHL
KATRIN BURMEISTER-LAMP
CATHERINE CLEOPHAS
NILS CRASSELT
BENEDIKT DOWNAR
KERSTIN FEHRE
MATTHIAS FINK
DAVID FLORYSIAK
GUNTHER FRIEDL
MARTIN FRIESL
FRANZ FUERST
WOLFGANG GÜTTEL
NINA KATRIN HANSEN
ANNE KATARINA HEIDER
CHRISTIAN HOFMANN
SVEN HÖRNER
STEPHAN KAISER
NADINE KAMMERLANDER
ALFRED KIESER
ALEKSANDRA KLEIN
NATALIA KLIEWER
DODO ZU KNYPHAUSEN-AUFSESS
SABINE T. KÖSZEGI
ARJAN KOZICA
CHRISTIAN KOZIOL
MARTIN KREEB
WERNER KUNZ
HANS-ULRICH KÜPPER
MICHAEL MEYER
JÜRGEN MÜHLBACHER
GORDON MÜLLER-SEITZ
J. PETER MURMANN
ANDREAS OSTERMAIER
BURKHARD PEDELL
ARTHUR POSCH
MARCEL PROKOPCZUK
TANJA RABL
SASCHA RAITHEL
NICOLE RATZINGER-SAKEL
ASTRID REICHEL
KATJA ROST
THOMAS RUSSACK
FLORIAN SAHLING
MARKO SARSTEDT
ANDREAS G. SCHERER
STEFAN SCHMID
UTE SCHMIEL
CHRISTIAN SCHMITZ
MARTIN SCHNEIDER
MARKUS SCHOLZ
LARS SCHWEIZER
DAVID SEIDL
THORSTEN SELLHORN
STEFAN SEURING
VIOLETTA SPLITTER
ANDREAS SUCHANEK
TILL TALAULICAR
ANN TANK
ANJA TUSCHKE
MATTHIAS UHL
CHRISTINE VALLASTER
PATRICK VELTE
CHRISTIAN VÖGTLIN
BARBARA E. WEISSENBERGER
ISABELL M. WELPE
HANNES WINNER
THOMAS WRONA
THOMAS ZWICK

JUNIOR MANAGEMENT SCIENCE

Johannes Witter, Predicting Stock Returns With Machine Learning: Global Versus Sector Models 561

Robin Roskosch, Beware of Bullshit – A Qualitative Study on Young Adults' Sustainability Awareness of Online Services 582

Nadhilla Mazaya, Board Gender Diversity: Evidence From Indonesia 609

Alexander Sake, Value Creation Opportunities of Generative AI – A Case Study 631

Justus Olbrich, The Effect of Changes in Internal Control Systems on Audit Risk 657

Jan Oliver Horstmann, Mandatory ESG Disclosure and Firm Value – A Quantitative Analysis of the Effect of Directive 2014/95/EU on Firm Value 677

Meret Anna Gläser, Government Interventions During the COVID-19 Pandemic, Culture, and Corporate Cost Behaviour 715

Zewei Shi, Modeling the Impact of Emission Credit Systems on Automotive Product Portfolios: A Mathematical Analysis of Policy Effects in Europe, China, and the U.S. Under Different Demand Scenarios 748

Hagen Alexander Hönerloh, Numerical Studies for the Scheduling of Continuous Annealing Lines 781

Lea Wedel, KPIs for Sustainability: Defining the Strategy for a Sustainable Future in the Insurance Industry 810

Published by Junior Management Science e.V.
This is an Open Access article distributed under the terms of the CC-BY-4.0 (Attribution 4.0 International). Open Access funding provided by ZBW.

ISSN: 2942-1861

Junior Management Science 10(3) (2025) 561-581



Junior Management Science

www.jums.academy
ISSN: 2942-1861



Predicting Stock Returns With Machine Learning: Global Versus Sector Models

Johannes Witter

Technical University of Munich

Abstract

Recent studies highlight the superior performance of non-linear machine learning models, such as neural networks, over traditional linear models in predicting cross-sectional stock returns. These models are capable of capturing complex non-linear interactions between predictive signals and future returns. This thesis researches whether sector-specific neural networks can detect sector-related relationships to outperform a global neural network. It evaluates the predictive power of these models at the stock level and in portfolios based on return forecasts, constructing long-short portfolios from the networks' sorted predictions. A global neural network model trained on the full sample of stocks dominates neural networks trained on individual GICS sectors in predicting the cross-section of US stock returns. Sector-specific neural networks fail to gain an advantage by capturing complex sector-specific interactions. They underperform the global neural network especially in the early out-of-sample period. The smaller sample size for each GICS sector requires a trade-off between model complexity and robust model estimation. Pooling the data for the global model solves this problem and supports the predictive power of neural networks for stock returns.

Keywords: cross-section of stock returns; machine learning; neural networks; return prediction; sector models

1. Introduction

I compare a global neural network with sector-specific neural networks to predict the cross-section of US stock returns. Therefore, I evaluate the predictive power of the two different models at the stock level and in portfolios constructed based on return forecasts. Recent research demonstrates the ability of non-linear machine learning models such as neural networks to outperform traditional linear models in predicting the cross-section of returns. I show how these neural networks perform better when trained on pooled data across sectors than on sector-specific data from GICS sectors. My data sample covers the sample period from July 1963 to December 2022, with an out-of-sample period from January 1994 to December 2022. The global neural network achieves a higher positive out-of-sample R^2_{OOS} than the sector models, which underperform a naive forecast of zero. For long-short portfolios based on the sorted predictions of the neural networks, the global model outperforms the sector models in terms of generated monthly returns and Sharpe ratio. The comparison of sector neural networks with simple OLS models highlights the necessary trade-off between estimating a stable model without overfitting and capturing complex sector-specific interactions.

I follow the machine learning training approach of Gu et al. (2020). I train neural networks with three hidden layers on 212 stock-specific signals from prior literature to predict the cross-section of returns for the US stock market. The neural networks are trained using a recursive scheme with increasing training samples and a fixed size rolling sample for validation. I refit all models once per year in December and predict monthly out-of-sample returns over the following year.

In total, I train eleven different neural network models. A global model is trained on the full sample of stocks. The

I would like to express my gratitude to all the people who have supported me in the realization of this thesis. In particular, I would like to thank my supervisor, Dr. Matthias Hanauer, for giving me the opportunity to work on this interesting topic and for his guidance and inspiration throughout the work on my thesis.

DOI: https://doi.org/10.5282/jums/v10i3pp561-581

© The Author(s) 2025. Published by Junior Management Science.
This is an Open Access article distributed under the terms of the CC-BY-4.0 (Attribution 4.0 International). Open Access funding provided by ZBW.

sector model consists of ten sector-specific neural networks trained on subsets of the sample data filtered for each of the ten sectors defined by the Global Industry Classification Standard (GICS). To assign stocks to sectors, I use existing data on GICS classifications and a custom mapping from the Standard Industrial Classification (SIC) system to GICS sectors.

First, I evaluate the two models on their the out-of-sample predictive performance for individual stock return forecasts. The global model outperforms the sector models in all individual GICS sectors and in the full sample. The global neural network produces a monthly R^2_{OOS} of 3.37% in the full sample and a positive R^2_{OOS} for each of the ten sectors. The sector models achieve a monthly R^2_{OOS} of -6.06% in the full sample, so they underperform a naive forecast of zero for all monthly stock returns. Sector models perform particularly poorly in predicting stock returns in sectors with small sample sizes.

To understand the differences in predictive performance, I examine the importance of each input variable in predicting returns with the neural network models. Variable importance is determined by the reduction in R^2_{OOS} that results from setting all values of a particular signal to zero while holding all other model estimates fixed. The different models share most of their most important variables. Analogous to Blitz et al. (2023), Size is the most influential signal in predicting returns in the global model and in most sector models. The neural networks tend to perform well in out-of-sample return forecasting when their relative importance is skewed towards Size. Sector models with a broader set of influential characteristics underperform out-of-sample.

I compare the profitability of portfolios based on the sorted predictions of the global model with the sector models. At the end of each month, I sort the stocks into decile portfolios and calculate the value-weighted returns of holding the decile portfolios over the next month. A long-short portfolio buys stocks with the highest expected returns and sells those with the lowest. Portfolio strategies based on the out-of-sample predictions of the global model outperform the sector-specific models. A long-short portfolio based on the global model's forecasts achieves an average monthly out-of-sample return of 2.71% and an annualized Sharpe ratio of 2.07. A long-short portfolio based on the forecasts of the sector models generates an out-of-sample monthly return of 0.99% and an annualized Sharpe ratio of 1.14. The outperformance of the global neural network is particularly strong in the early years of the out-of-sample period. During this period, the global model achieves its highest returns while the sector models struggle to remain profitable.

The sector-specific neural networks continue to underperform when compared to simple ordinary least squares (OLS) models. A long-short portfolio based on the sorted predictions of the OLS sector models generates a higher value-weighted return than the sector neural networks. However, the OLS outperformance comes only from the first half of the out-of-sample period when less training data is available. Small sector sample sizes require a trade-off between stable model estimation and capturing complex sector-specific interactions. The global neural network trained on pooled data significantly outperforms the OLS models.

The global neural network shows some out-of-sample sector allocation power. In the cross-section of sectors, it correctly predicts higher relative returns for the most profitable sectors and lower returns for the least profitable sectors. As a result, the returns generated by the global model are lower with sector-neutral portfolios.

Section 2 reviews the recent literature on machine learning for return forecasting and global versus industry-specific models. Section 3 presents the sources of stock data and input signals for return prediction and explains the sector classifications. Section 4 describes the methodology used to train neural networks for return prediction and to construct portfolios based on these predictions. Section 5 presents the results of comparing the forecasting performance of the global model with the sector models. Section 6 concludes.

2. Literature review

The past decades produced a variety of literature focusing on predicting the cross-section of stock returns. Authors explore a variety of variables in linear models, but there is still a lack of consensus regarding which variables are related to expected stock returns. This problem is often referred to as factor zoo. Linear models cannot deal with many variables and their potential nonlinearities and interactions. Therefore, recent research focuses on more complex machine learning models to handle the high dimensionality in the factor zoo.

Early literature focuses on single machine learning models and their ability to outperform traditional methods. Moritz and Zimmermann (2016) propose tree-based conditional portfolio sorts as a machine learning approach. In their models, recent past returns within the last six months predict future returns and outperform linear models like Fama-MacBeth regressions. Excess returns persist even after accounting for transaction costs and common risk factors. Traditional methodologies with linear assumptions fail to capture a nonlinear relationship between past and future returns.

A nonparametric model using adaptive group least absolute shrinkage and selection operation (LASSO) isolates relevant predictors in a high-dimensional setting in Freyberger et al. (2020). A small subset of variables, including size, total volatility, and recent return-based metrics, provide unique predictive power. Their model significantly outperforms linear approaches like those of Lewellen (2015) with higher out-of-sample Sharpe ratios. The nonparametric model selects fewer variables in-sample than the linear models but captures nonlinear interactions.

Gu et al. (2020) are among the first to present a comparative analysis of machine learning models for predicting stock returns. Their models agree on a small set of dominant variables, with price trends, liquidity, and volatility as the most influential predictors. Neural networks perform the best among all machine learning models, and portfolios sorted on neural network return predictions double the

Sharpe ratios of linear models. Shallow neural networks outperform deeper ones due to the limited data and low signal-to-noise ratio in empirical asset pricing. Interactions and nonlinear relations between variables drive the outperformance of machine learning methods. Azevedo and Hoegner (2023) report similar results with tree-based and neural network approaches. Their machine learning methods uncover interaction effects challenging traditional risk-based explanations in asset pricing. Neural network models utilize 299 stock market anomalies to achieve monthly out-of-sample returns 1% higher than a linear benchmark. Linear regressions are easy to interpret but underperform their machine learning models using statistical significances and returns.

Most recent literature uses technically more advanced machine learning strategies. Azevedo et al. (2024) examine the expected returns of deep learning strategies. Their long short-term memory (LSTM) models yield net returns of up to 1.42% per month, even after accounting for the recent era of high liquidity, transaction costs, and post-publication decay. Strategies combining several machine learning models constantly achieve significant returns after cost. Cost mitigation techniques reduce turnover and trading costs but do not improve net performance. L. Chen et al. (2024) combine three deep neural network models with no-arbitrage constraints to estimate asset pricing models for US stocks. Incorporating specific domain knowledge into the technical implementation enhances prediction accuracy and out-of-sample performance. Their deep learning strategies with no-arbitrage constraints outperform other machine learning benchmarks in Sharpe ratio and identify the core variables driving asset prices.

Other research focuses on using machine learning to predict stock returns not only in the US but globally. Tobek and Hronec (2021) aggregate 153 anomalies across global markets into one mispricing signal using machine learning. Their strategy outperforms linear models out-of-sample in various international markets. Extending the training sample with international data does not improve out-of-sample performance for the US market. However, machine learning models trained on US stocks perform well in markets outside the US. Cakici et al. (2023) investigate machine learning's cross-sectional return predictability across 46 global stock markets. Combining predictions from multiple machine learning models delivers robust out-of-sample returns across diverse markets. Developed markets show higher profitability than emerging ones. Firm size and idiosyncratic risk are the most important variables for predictions, with higher returns in smaller firms and markets with more idiosyncratic risk. Azevedo et al. (2023) also focus on the out-of-sample performance of different machine learning models across an international data sample. Neural networks and composite predictors perform the best. These models achieve significant monthly long-short returns of around 2%. Portfolio returns remain significant even after transaction costs and outperform linear benchmark models. Drobetz and Otto (2021) use machine learning strategies to predict European stock returns. Like in the US market, machine learning models outperform traditional linear models by capturing nonlinearities and variable interactions. Neural networks and classification-based approaches perform best and generate significant returns even after transaction costs. Support vector machines, which classify stocks into decile portfolios, deliver even higher returns by eliminating the noise of expected returns at the stock level. Leippold et al. (2022) apply machine learning models to the Chinese stock market. In a market dominated by retail investors, liquidity and volatility indicators have predictive power over traditional variables like valuation ratios. Neural networks perform best, particularly for small-cap and non-state-owned firms. They achieve higher predictability in China than in the US due to distinct asset pricing dynamics driven by local investor behavior. Hanauer and Kalsbach (2023) assess various machine learning models for predicting stock returns in a broad sample of emerging markets. Like in developed markets, their models identify nonlinearities and interactions among variables. Tree-based methods and neural networks deliver superior long-short returns and alphas over linear models. Efficient trading rules ensure machine learning predictions outperform even after transaction costs, short-selling constraints, and limiting the sample to big stocks.

Despite the strong performance of machine learning models for stock market prediction, there are still problems in implementing them in practice. Rasekhschaffe and Jones (2019) focus on mitigating overfitting in machine learning models. Feature engineering and forecast combinations decrease the risk of overfitting. These techniques increase the signal-to-noise ratio and produce more robust predictions. Avramov et al. (2023) criticize high limits-to-arbitrage environments and exclude stocks like microcaps and distressed firms. Machine learning portfolios often rely on long and short positions that are impossible in practice. The profitability of machine learning strategies is reduced when trading costs are considered due to high turnover. The authors propose including trading costs in machine learning models and imposing economic restrictions. Blitz et al. (2023) report the impact of varying prediction horizons in machine learning models for stock return predictions. While one-month forecasts yield high gross returns, the net returns considering transaction costs are minimal after 2004. Machine learning models with longer prediction horizons provide significant net alpha due to reduced portfolio turnover. One-month forecasts rely on short-term price signals, whereas longer horizon predictions rely more on value-oriented signals. Aligning the design of machine learning models with trading horizons enhances profitability through reduced turnover and better after-cost performance.

In this thesis, I focus on sector-specific versus global machine learning models. Therefore, it is appropriate to consider prior research on the relation between industry-specific and market predictions. This includes traditional methods of predicting the cross-section of stock returns, like factor investing and other linear regression models. Kim et al. (2013) enhance the linear residual income model (RIM) with industry-specific factors using the value-to-book (V/B)

ratio. Decomposing V/B into industry-specific components better predicts future abnormal returns and outperforms traditional RIM implementations. Their industry-adjusted RIM model provides superior predictive accuracy for abnormal returns compared to conventional valuation measures. Liu et al. (2014) analyze the predictive power of industry effects in option-implied volatility measures on stock returns. Industry-neutral portfolios outperform full-universe portfolios with higher Sharpe ratios and lower downside risk. Cavaglia et al. (2006) explore region-neutral and industry-neutral value portfolios. Industry-neutral portfolios offer more stable returns and higher information ratios due to lower volatility and less cyclicality. They capture the global value premium more effectively and provide a better risk-return profile.

Modern machine learning models also capture economically meaningful interdependencies among industries. Rapach et al. (2019) examine cross-industry return predictability with machine learning models like LASSO. Due to gradual information diffusion, lagged returns from industries like financial and commodity sectors can predict returns in other sectors. An industry-rotation strategy based on these cross-industry signals outperforms linear methods. Ciner (2019) uses a random forest model as a machine learning strategy to predict market returns with industry returns. Industry returns provide significant out-of-sample predictive power for the market index. Random forests outperform traditional linear models due to their capacity to capture both linear and non-linear dynamics. Industry-level information forecasts market movements in a way that linear models fail to capture.

This thesis extends the existing literature on machine learning in empirical asset pricing by integrating it with industry-specific strategies. Building upon the work of Gu et al. (2020), I use neural networks as well-performing machine learning models to forecast the cross-section of stock returns for the US market. I compare long-short portfolio returns achieved by sector-specific models with those of a global model.

3. Data

3.1. Stock data

My sample includes all US stocks listed on the NYSE, AMEX, and NASDAQ. The sample period runs from July 1963 to December 2022. I source equity returns and other stock market data from CRSP. Accounting data to replicate the Fama and French (2015) five-factor model is from Compustat.

I calculate monthly excess returns as the one-month stock return from CRSP over the risk-free rate provided on Kenneth R. French's homepage.[1] To predict the cross-section of stock returns, I subtract the monthly cross-sectional median excess return from the monthly excess return for each stock. "Returns" throughout this thesis denote these relative monthly stock returns with the market component already removed.

The input variables to the machine learning models come from A. Y. Chen and Zimmermann (2022). I download the August 2023 version of signals from their Open Source Asset Pricing (OSAP) website.[2] This includes 209 firm-level characteristics replicated from the academic asset pricing literature. I use the terms variables, signals, and characteristics interchangeably throughout this thesis.

In addition to the 209 signals from the OSAP data, I define three input variables from CRSP data. Short-term reversal is the prior one-month return, Price is the natural logarithm of the CRSP price data field, and Size is the natural logarithm of the price multiplied by the shares outstanding.

All input variables are signed so that higher values correspond to higher expected returns. Following Gu et al. (2020), Blitz et al. (2023) and other recent literature, I rank all input variables cross-sectionally for each month into the interval of [-1,1]. This helps neural network models deal with varying ranges of values and different variances across the signals during training. Missing values are filled with the monthly cross-sectional median rank.

I follow Hou et al. (2020) and Blitz et al. (2023) and exclude microcaps from my sample to prevent them from driving my results. I define microcaps as all stocks with a monthly market capitalization below the 20th percentile of the NYSE market capitalization in that month.

After excluding microcaps, my full sample from July 1963 to December 2022 includes approximately 1.3 million monthly stock observations with a monthly average of 1842 stocks.

3.2. Sector data

To train sector-specific machine learning models, I assign all stocks in my sample to sectors according to the Global Industry Classification Standard (GICS) from MSCI and S&P. I categorize stocks into ten sectors defined by GICS: Energy, Materials, Industrials, Consumer Discretionary, Consumer Staples, Health Care, Financials, Information Technology, Communication Services, and Utilities. Due to the small number of observations, I include the Real Estate sector in the Financials sector. This corresponds to the GICS classification before 2016, better representing the biggest part of my sample period.

I prefer the GICS over the Standard Industrial Classification (SIC) system often used in other literature on industry specifics. The GICS is the more modern industry taxonomy with a stronger focus on new industries like the computer, software, and information technology sectors.

I source data on SIC industry classification from CRSP and data on GICS sector classification from Compustat. GICS data has weaker coverage than SIC data, particularly from

[1] See http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html (2024).

[2] See https://www.openassetpricing.com/august-2023-data-release/ (2023).

the beginning of my sample until 1985. Therefore, I develop a mapping from SIC industries to GICS sectors described in Table 1. If a monthly stock observation is not assigned to a sector under GICS but to a SIC industry, I fill in the missing GICS classification based on the mapping in Table 1.

The mapping is based on two different inputs. First, I consider overlaps between the SIC and GICS classifications for stock observations with both data points in my sample. For example, 90% of all stocks classified as Financials in SIC are also classified as Financials in GICS. Similarly, over 80% of all stocks classified as utilities in SIC are also classified as Utilities in GICS. Second, I consider the definitions of the industry taxonomy in both classification systems.[3] For example, SIC includes transportation, communications, electric, gas, and sanitary services in one Division. GICS classifies Transportation as an industry group in the sector Industrials, communications belong to the sector Communication Services, and electric, gas, and sanitary services belong to the sector Utilities. This enables me to map the corresponding SIC major groups to the GICS sectors.

Table 1 includes the ten GICS sectors used for sector-specific machine learning models. It summarizes all SIC digit codes and the corresponding industry description mapped to each GICS sector. For example, I map the SIC codes 1200-1399 (Major Group Coal Mining and Major Group Oil And Gas Extraction) and 2900-2999 (Major Group Petroleum Refining And Related Industries) to the GICS sector Energy. The SIC industries mapped to a GICS sector can be either entire Divisions (determined by a capital letter) or more detailed Major Groups (determined by the first two digits) and Industry Groups (determined by the first three digits).

If I recognize no clear relationship between a SIC industry and GICS sectors, then I don't include this SIC industry in the mapping. A stock with missing GICS data and a SIC code not contained in the mapping will not be assigned to a GICS sector. The same happens to stocks with missing data for both classification systems. After applying the mapping, 15,239 out of 1.3 million monthly stock observations in the full sample are not assigned to a GICS sector. They are included in the training data for the global neural network but not in the training data for the sector-specific neural networks.

4. Methodology

4.1. Return prediction using machine learning

My methodology follows Gu et al. (2020) and Hanauer and Kalsbach (2023), with the difference that I train one global model and one sector-specific model for each of the ten GICS sectors.

I aim to predict the cross-section of US stock returns, so I forecast the outperformance of a stock relative to the US stock market. The relative return of a stock is defined as

$$r_{i,t}^{rel} = r_{i,t} - Mkt_t, \tag{1}$$

where $r_{i,t}$ is the excess return of stock i in month t and Mkt_t is the cross-sectional median excess return across all stocks in the sample in month t.

I describe the one-month-ahead relative return of a stock $r_{i,t+1}^{rel}$ as an additive prediction error model:

$$r_{i,t+1}^{rel} = E_t\left[r_{i,t+1}^{rel} \mid x_{i,t}\right] + \epsilon_{i,t+1}. \tag{2}$$

$E_t[r_{i,t+1}^{rel}|x_{i,t}]$ is the conditional expected relative return of stock i in month t for month $t+1$. It is conditional as it depends on $x_{i,t} \in \mathbb{R}^p$, a vector of stock-specific p input variables known at month t. $\epsilon_{i,t+1}$ is the prediction error term.

I estimate the expected relative return with the unknown function $f^*, f^* : \mathbb{R}^p \to \mathbb{R}$. It estimates the expected return depending only on the vector of p stock-specific input variables available in month t:

$$E_t\left[r_{i,t+1}^{rel} \mid x_{i,t}\right] = f^*(x_{i,t}). \tag{3}$$

In the case of neural networks, the unknown function $f^*(x)$ is approximated by a nonlinear function $f(x, \theta, \rho)$. This function is parametrized by a vector of coefficients θ and a set of hyperparameters ρ. When training neural networks, the coefficients θ are estimated from the training data with respect to the hyperparameters ρ and a predefined loss function L. The hyperparameters ρ are optimized concerning the loss function L based on the estimated coefficients θ and available data.

Neural networks as a form of supervised machine learning outperform linear models in prior literature (Azevedo & Hoegner, 2023; Gu et al., 2020). Therefore, I choose three-layer neural networks as the machine learning model for this thesis. Appendix A4 describes the model architecture and hyperparameters used to train the neural networks. In addition, I later use ordinary least squares (OLS) models as a benchmark for sector-specific neural networks.

The global neural network model takes the full sample of 1.3 million monthly stock observations as input. The sector-specific machine learning models for the ten GICS sectors take all monthly stock observations assigned to the respective GICS sector as input. Therefore, the samples used to train the sector-specific neural networks differ significantly in their number of observations, depending on the sector's size.

To avoid data leakage, I divide all input samples into three disjoint time periods, which always keep the temporal ordering of the data: the training, validation, and testing samples. First, I estimate the neural network coefficients for a range of hyperparameter values on the training sample. The validation sample compares the loss function results for each set of hyperparameters based on the estimated model from the

[3] See https://www.msci.com/our-solutions/indexes/gics for GICS definitions and https://www.osha.gov/data/sic-manual for SIC definitions.

Table 1: Mapping of SIC to GICS industries

This table maps industries classified under the Standard Industrial Classification (SIC) system to sectors classified under the Global Industry Classification Standard (GICS) system used in this thesis. The first column contains the 11 different sectors of the GICS system. Throughout this thesis, the 11th sector, Real Estate, is not considered separately, but is included in the Financials sector. The second and third columns contain the SIC industries mapped to the GICS sector in the first column. The second column contains the SIC digit code, and the third includes the corresponding industry description. The SIC industries mapped to a GICS sector can be either entire Divisions (determined by a capital letter) or more detailed Major Groups (determined by the first two digits) and Industry Groups (determined by the first three digits). This mapping is used to classify individual stocks into sectors when GICS sector information is unavailable in Compustat. If a SIC industry classification is available, the GICS sector is added according to this mapping. If no SIC classification is available either, the GICS sector value is 'Missing', and the stock is not included in the training data for the sector models.

GICS Sector	SIC Code	SIC Description
10 - Energy	1200–1399	Coal Mining and Oil/Gas Extraction
	2900–2999	Petroleum Refining and Related Industries
15 - Materials	Division B 1000–1499 (excluding 1200–1399)	Mining (excluding Coal Mining and Oil/Gas Extraction)
	2400–2499	Lumber and Wood Products, Except Furniture
	2600–2699	Paper and Allied Products
	3300–3399	Primary Metal Industries
20 - Industrials	Division C 1500–1799	Construction
	Division E 4000–4999 (excluding 4800–4999)	Transportation (excluding Communications and Utilities)
	Division J 9100–9999 (excluding 9900–9999)	Public Administration (excluding Nonclassifiable Establishments)
	3400–3499	Fabricated Metal Products, Except Machinery and Transportation Equipment
	3500–3599	Industrial and Commercial Machinery
	7320–7329	Credit Reporting and Collection
	7340–7349	Services to Dwellings and Other Buildings
	7360–7369	Personnel Supply Services
	7390–7399	Miscellaneous Business Services
	7500–7599	Automotive Repair Services and Parking
	7600–7699	Miscellaneous Repair Services
	8710–8719	Engineering Architectural and Surveying Services
	8740–8749	Management and Public Relations
	8900–8999	Services Not Elsewhere Classified
25 - Consumer Discretionary	Division G 5200–5999 (excluding 5400–5499)	Retail Trade (excluding Food Stores)
	Division F 5000–5199 (excluding 5140–5189, 5180–5189)	Wholesale Trade (excluding Groceries and Beer, Wine, and Distilled Alcoholic Beverages)
	1500–1599	Building Construction General Contractors and Operative Builders
	2200–2299	Textile Mill Products
	2300–2399	Apparel and Other Finished Products Made from Fabrics and Similar Materials
	2500–2599	Furniture and Fixtures
	3100–3199	Leather and Leather Products

Table 1 — continued

GICS Sector	SIC Code	SIC Description
25 - Consumer Discretionary	3900–3999	Miscellaneous Manufacturing Industries
	7000–7099	Hotels, Rooming Houses, Camps, and Other Lodging Places
	7200–7299	Personal Services
	7800–7899	Motion Pictures
	7900–7999	Amusement and Recreation Services
30 - Consumer Staples	Division A 0100–0999	Agriculture, Forestry, and Fishing
	2000–2199	Food Products and Tobacco Products
	5140–5159	Wholesale Trade - Groceries
	5180–5189	Wholesale Trade - Beer, Wine, and Distilled Alcoholic Beverages
	5400–5499	Food Stores
35 - Health Care	2800–2899	Chemicals and Allied Products (including Drugs)
	3840–3849	Surgical Medical and Dental Instruments and Supplies
	3850–3859	Ophthalmic Goods
	8000–8099	Health Services
	8300–8399	Social Services
	8730–8739	Research Development and Testing Services
	9900–9999	Nonclassifiable Establishments
40 - Financials	Division H 6000–6799	Finance, Insurance, and Real Estate
45 - Information Technology	3570–3579	Computer and Office Equipment
	3600–3699	Electronic and Other Electrical Equipment and Components
	3820–3829	Laboratory Apparatus and Analytical Optical Measuring and Controlling Instruments
	7370–7379	Computer Programming Data Processing
50 - Communication Services	4800–4899	Communications
55 - Utilities	4900–4999	Electric, Gas, and Sanitary Services
60 - Real Estate	Included in GICS sector 40 - Financials	

training sample. The optimal hyperparameter set minimizes the loss function on the validation sample and is then used to retrain five different neural networks on the training sample. I use these five models to predict the monthly returns for the test sample. The final prediction for each stock is the average over the five individual model predictions to reduce the variance in single forecasts.

Following the training approach as in Gu et al. (2020) and Blitz et al. (2023), I retrain the models once at the end of every year but predict every month using the latest model and data. The first 18 years of my sample (July 1963 to December 1981) are the first training sample, and the next 12 years (January 1982 to December 1993) the first validation sample. The first one-year test sample is the following 12 months, so the first out-of-sample (OOS) prediction is made for January 1994. To predict the monthly returns from January 1995 to December 1995, I extend the training sample by one year (July 1963 to December 1982) and roll forward the validation sample by one year (January 1983 to December 1994). I repeat this procedure for each year in my sample. No future information is leaked from a previous period.

To evaluate the predictive performance for individual stock return forecasts on the test sample, I use the pooled out-of-sample R^2_{OOS} defined by Gu et al. (2020):

$$R^2_{OOS} = 1 - \frac{\sum_t^T \sum_i^N \left(r^{rel}_{i,t} - \hat{r}^{rel}_{i,t}\right)^2}{\sum_t^T \sum_i^N \left(r^{rel}_{i,t}\right)^2}. \tag{4}$$

This metric compares the out-of-sample forecasts with a naive forecast of zero, better suited to individual return predictions than the typical forecast with mean returns.

4.2. Variable importance

As a primary measure to interpret the results of the machine learning models, I rank the respective input variables according to their variable importance. This aims to identify characteristics that influence the cross-section of expected returns. Following Gu et al. (2020) and Blitz et al. (2023), I define variable importance as the reduction in panel predictive out-of-sample R^2_{OOS}. For each annually trained neural network, I iteratively set the values for each input variable to zero while holding the model estimates fixed. In each iteration, I predict new monthly returns for the respective one-year test sample and calculate the change in out-of-sample R^2_{OOS}. I use the average variable importance across each annually trained model to rank each feature where rank one is the most important characteristic. To determine the relative importance of individual variables for the performance of each model, I normalize variable importance within a model to sum to one. If setting a signal to zero increases the panel predictive out-of-sample R^2_{OOS}, the variable importance measure for that signal is negative. Therefore, the normalized variable importance of other signals within a model can be greater than 1.

4.3. Machine learning portfolios

Portfolio performance is my primary metric for evaluating the forecast performance of machine learning models. At the end of each month, each model produces a prediction of a stock's next month's relative return $\hat{r}^{rel}_{i,t+1}$. Based on these forecasts, I sort stocks from highest to lowest predicted return and assign them into decile portfolios using NYSE breakpoints.[4] I reassign and rebalance portfolios at the end of each month. I compute value-weighted returns from holding the decile portfolios over the next month to avoid small stocks driving the results. Finally, I construct a zero-net investment (long-short) portfolio that goes long in the highest decile portfolio (decile 10) and short in the lowest decile portfolio (decile 1).

I evaluate the predictive performance of three different machine learning strategies. The first strategy forms decile portfolios based on the predictions of the global neural network model. The second strategy forms decile portfolios with sector-neutral portfolio sorts based on the global neural network model predictions. This means each sector gets individual breakpoints for the decile sorts. The third strategy forms decile portfolios based on the predictions from the sector-specific neural networks. First, I perform portfolio sorts for each sector individually based on the respective model forecasts. Then, I combine the sector-specific decile portfolios into single decile portfolios. For example, the top decile portfolio for month t contains all stocks from the ten top decile portfolios across all sectors in month t.

[4] Breakpoints for the decile sorts are first determined using only stocks listed on the NYSE. All stocks are then sorted into decile portfolios based on these breakpoints, regardless of which exchange they are listed on. As the NYSE contains stocks with larger average market capitalizations, this reduces the influence of small stocks on the portfolio sorts.

To compare the results of the three neural network portfolio sorts, I provide each decile portfolio's average predicted returns, realized returns, and Sharpe ratios. I compute mean returns and associated t-statistics for the long-short portfolios of each machine learning strategy. To benchmark the long-short returns, I consider the adjusted R^2-value and alphas from the Capital Asset Pricing Model (CAPM) and Fama and French (2015) five-factor model with their associated Newey and West (1987) adjusted t-statistics using six lags. The factors are based on the same sample of 1.3 million monthly stock observations as the neural network portfolios.

Finally, I benchmark the predictive performance of the sector-specific neural networks with OLS models in Section 5.4. The portfolio sorts work similarly but are based on predictions from sector-specific OLS models. All sector models mentioned outside of Section 5.4 always refer to sector-specific neural networks.

5. Empirical results

5.1. Prediction performance

The global neural network dominates the sector-specific neural networks in out-of-sample predictive performance for individual stock return forecasts. On the full sample over the out-of-sample period from January 1994 to December 2022, the global model achieves a monthly R^2_{OOS} of 3.37%. The sector models achieve a monthly R^2_{OOS} of -6.06%, so they underperform a naive forecast of zero for all monthly stock returns. Across all ten GICS sectors, the global model outperforms the respective sector model. Sector models perform particularly worse for sectors with only a small sample size.

Table 2 compares the monthly out-of-sample stock-level prediction performance across all ten GICS sectors between the global model and the ten sector-specific models. The sector performance for the global model is determined by filtering the out-of-sample predictions of the global neural network for the respective sector stocks. In addition, Table 2 includes the average monthly observations per sector over the full sample period from July 1963 to December 2022. This demonstrates the sample size for each sector.

The global model produces positive R^2_{OOS} statistics across all individual sectors. Taking the full sample of stocks, the global model achieves a R^2_{OOS} of 3.37%. Therefore, the predictions consistently outperform a naive forecast of zero to all stocks in all months over the out-of-sample period. Except for Utilities, the R^2_{OOS} statistics for all sectors are larger than 2%. Over six out of ten sectors the global model produces R^2_{OOS} above 3%, with the highest value at 4.57% for the sector Health Care. Utilities appear to be an outlier with 0.11%, more than an order of magnitude smaller than the R^2_{OOS} for all other sectors. There is no correlation between the sectors' sample size and the predictive performance of the global neural network. The model achieves a R^2_{OOS} of 3.11% on Communication Services, the sector with the smallest sample size and only an average of 58 stocks per month in the sample. This is more than the 2.26% for Financials, the biggest sector

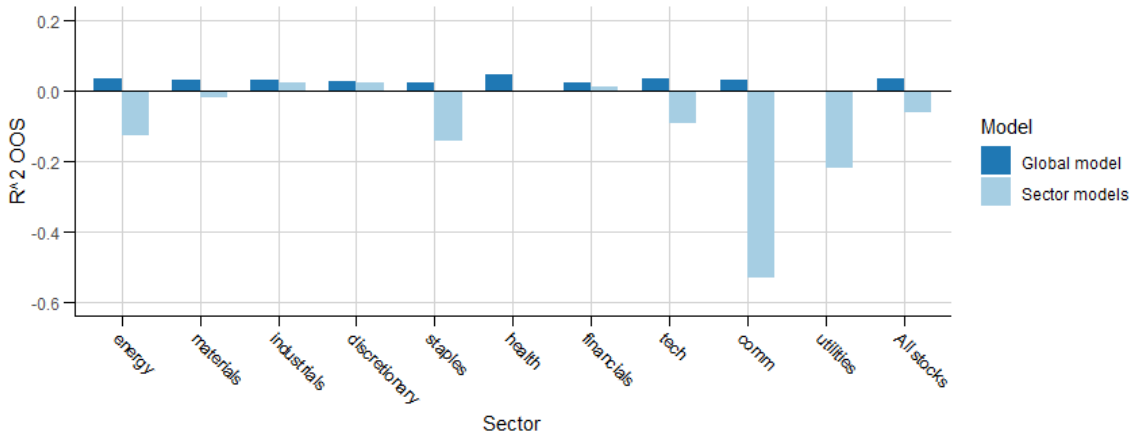
Table 2: Monthly out-of-sample stock-level prediction performance

This table summarizes the monthly out-of-sample stock-level prediction performance across all GICS sectors using the global model and the respective neural network sector models. Panel A reports the monthly R^2_{OOS} statistics of the global model and the respective neural network sector models for the full sample over the out-of-sample period from January 1994 to December 2022. The third column in Panel A reports the average monthly stocks per sector in the full sample over the full sample period from July 1963 to December 2022. The sample consists of US CRSP stocks, excluding microcap stocks with a market capitalization smaller than the 20th percentile of stocks listed on the NYSE. Panel B provides a visual comparison of the monthly R^2_{OOS} statistics in Panel A.

Panel A: Percentage R^2_{OOS}

Sector	Global Model	Sector Models	Avg. monthly observations
Energy	3.49	-12.79	107
Materials	3.23	-1.92	136
Industrials	3.07	2.21	293
Consumer Discretionary	2.86	2.17	286
Consumer Staples	2.47	-14.11	113
Health Care	4.57	-0.36	184
Financials	2.26	1.18	306
Information Technology	3.43	-9.17	223
Communication Services	3.11	-53.07	58
Utilities	0.11	-21.85	113
All stocks	3.37	-6.06	1842

Panel B: Visual comparison of R^2_{OOS} statistics



with a mean of 306 monthly stock observations in the sample.

The sector-specific neural networks perform significantly worse than the global model across the whole sample of stocks. Sector models produce negative R^2_{OOS} in seven out of ten sectors. The R^2_{OOS} statistic on the full sample is also negative with -6.06%. This means a naive forecast of zero to all stocks in all months dominates the sector models over the full sample. The sector neural networks achieve positive R^2_{OOS} only in the sectors Industrials, Consumer Discretionary and Financials. These are the three biggest sectors by sample size, with around 300 stock observations per month. The models perform poorly in all sectors with only a small sample size. Sectors like Energy, Consumer Staples, Communication Services, and Utilities are the smallest sectors, with monthly stock observations in the range from 58 to 113, and all produce two-digit negative R^2_{OOS}. Communication Services has only half of the sample size compared to the next bigger sector, and the respective sector model performs by far the worst with a R^2_{OOS} of -53.07%. Information Technology is an exception with 223 average monthly stock observations and a negative R^2_{OOS} of -9.17

Based on the correlation between sector size and sector model performance, complex nonlinear machine learning models like neural networks might not be suited for small samples. The number of coefficients estimated in a neural network rapidly expands with 212 input signals. Therefore, the observations-to-parameters ratio is particularly small in sectors with a small sample size. Pooling all sectors as input for the global neural network improves the observations-to-

parameters ratio. This motivates benchmarking the sector neural networks against sector OLS models in Section 5.4.

Panel B of Table 2 visualizes the negative outliers in the predictive performance of the sector models. The global neural network achieves a constant positive R^2_{OOS} in a similar range over the whole sample. The sector neural networks never dominate the global model and produce apparent downward deviations that stand out from the range of other values.

5.2. Variable importance

The neural network models with the best out-of-sample prediction performance share most of their most important variables. Size is the most influential signal in predicting returns in the global model and across most sector models. Neural networks tend to perform well in out-of-sample return forecasts if their relative importance is skewed towards Size. Risk measures, liquidity variables, and recent price trends are other important characteristics in well-performing models. Sector models with a broader set of influential characteristics underperform out-of-sample.

I examine the importance of individual input signals in predicting returns with neural network models. To determine variable importance, I calculate the reduction in R^2_{OOS} from setting all values of a particular signal to zero while keeping all other model estimates fixed. This process is repeated iteratively over all variables for each model. The average over all annual test samples results in a single importance measure per model for each variable.

Figure 1 visualizes each neural network model's variable importance of the top 20 stock-level signals. This includes the global model and sector-specific models for each of the ten GICS sectors. Variable importances within a model are normalized to sum to one, giving them the interpretation of relative importance for that model.

Figure 2 reports overall rankings of variable importance across the 100 most influential characteristics for all neural network models. I rank the importance of individual signals for each model (with rank one as the most important) and then sum up their ranks. Variables in Figure 2 are ordered so that the highest total ranks are at the top, and the lowest-ranking characteristics are at the bottom. A darker color in a column indicates a higher importance of the respective variable for the individual model.

The most essential variables are similar among many of the neural networks. Analogous to Blitz et al. (2023), Size is the most influential variable overall. It is first in overall ranks and has the highest relative importance for the predictive performance of nine models. It dominates all other variables by a large margin. Except for Consumer Staples (*staples*), the relative importance is highly skewed towards Size in all models where it is the most influential variable. Sector models for Communication Services (*comm*) and Utilities (*utilities*) are the only ones without Size as the most important signal. They are more democratic and draw information from a broader set of predictive variables. However, they underperform all other models in predictive performance, as seen in Table 2. Neural networks tend to perform better out-of-sample in Table 2 when the relative importance in Figure 1 is skewed towards Size.

The best-performing sector models like Industrials (*industrials*), Consumer Discretionary (*discretionary*), and Financials (*financials*) focus on the same set of influential variables as the global model (global). Besides the most important signal Size, this includes Idiosyncratic risk (*IdioVolAHT, IdioVol3F*) and Realized Volatility (*RealizedVol*) as risk measures; Amihud (2002) illiquidity (*Illiquidity*) and Volume Variance (*VolSD*) as liquidity variables; Short-term Reversal (*STreversal*) and Trand Factor (*TrendFactor*) as recent price trends; and to a lesser extent valuation ratios and fundamental signals like earnings-to-price (*EP*) and sales-to-price (*SP*).

Looking at Figure 2, the highest-ranked characteristics are the ones that achieve solid relative importance across all models. The four top-ranked signals are Size, Institutional Ownership and Idiosyncratic Volatility from Nagel (2005) (*RIO_Volatility*), Short-term Reversal, and Price. They are among the top-ranked variables in many individual models. However, the following top 25 ranked characteristics do not belong to top-ranked variables in any particular model but instead stay at least relatively important throughout all models. Forecasting power for the rest of the 100 most influential variables in Figure 2 is more heterogeneously distributed. The signals achieve high importance for the predictive power of some individual models but are not relevant in others. For example, risk measures like Idiosyncratic risk and Realized Volatility are among the top signals in the global model and parts of the sector models but irrelevant to the R^2_{OOS} of other sector models.

5.3. Neural network portfolios

Portfolio strategies based on the global model's sorted out-of-sample predictions outperform the sector-specific models. A long-short portfolio based on the global model's forecasts achieves an average monthly out-of-sample return of 2.71% and an annualized Sharpe ratio of 2.07. The portfolio strategy is profitable throughout the out-of-sample period and generates significant alphas over established factor models. The outperformance of the global neural network is particularly strong from the beginning of the out-of-sample period through to the early 2000s. The sector neural networks predict very extreme portfolio returns but achieve significantly lower returns than the global model. A long-short portfolio based on the forecasts of the sector models has an out-of-sample monthly return of 0.99% and an annualized Sharpe ratio of 1.14. The portfolio strategy generates small alphas on top of the factor models but performs unreliably in the first half of the out-of-sample period and only starts to recover thereafter.

After comparing the predictive ability of the global model versus the sector models for individual stock returns in Section 5.1, I assess the profitability of portfolios based on the sorted predictions from my neural networks trained to forecast out-of-sample returns. Each neural network produces

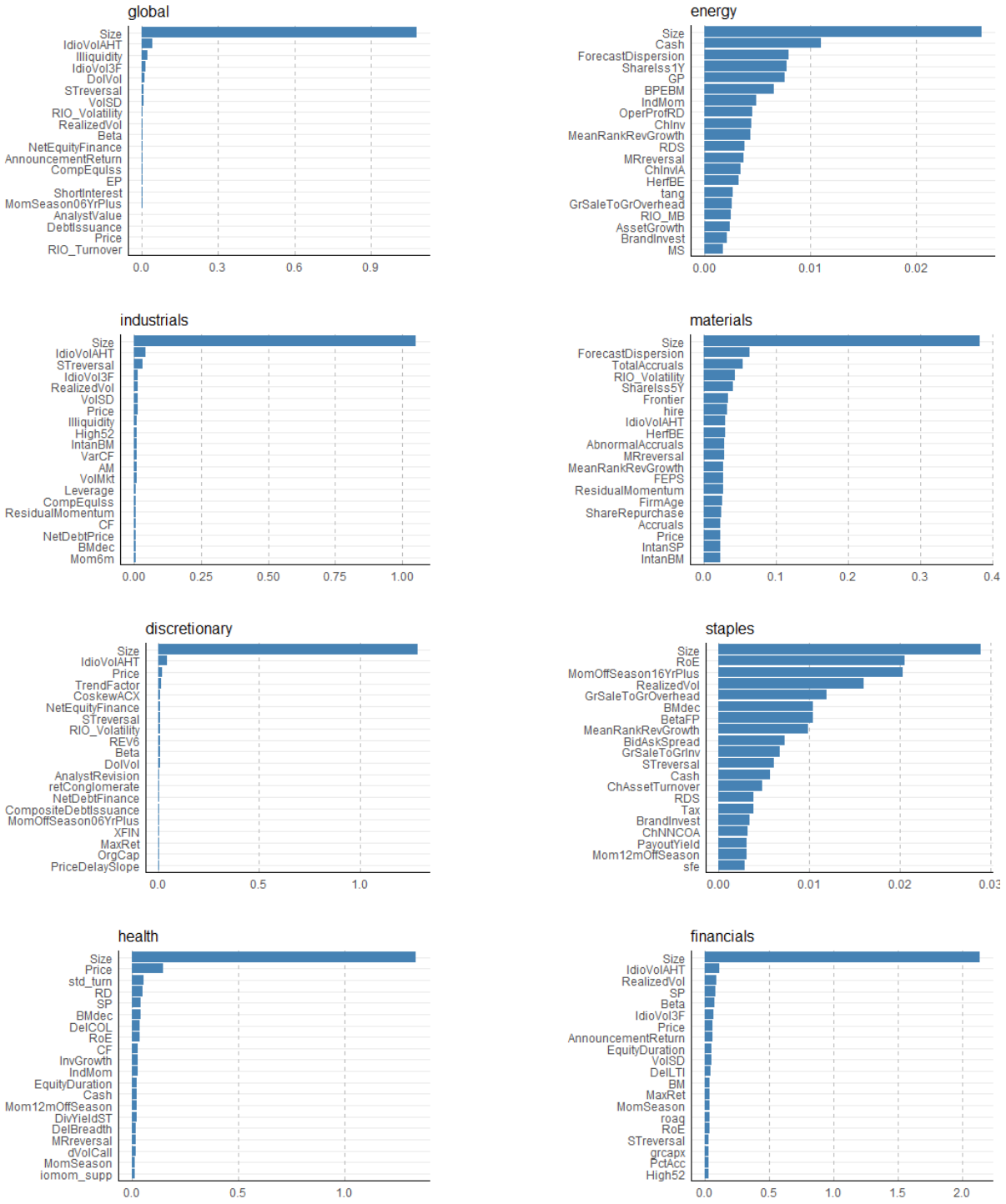
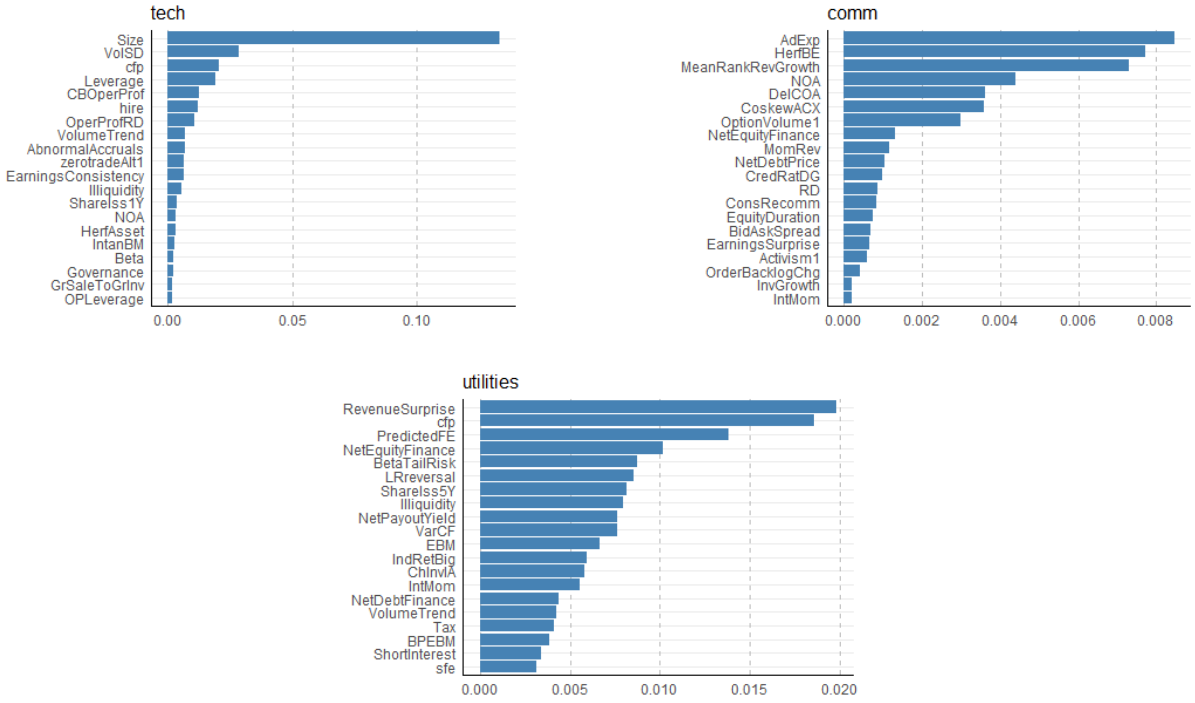


Figure 1: Variable importance by model

This figure plots the variable importance for each model's top 20 most influential characteristics. This includes the global model (*global*) and all ten sector models (GICS sectors Energy (*energy*), Materials (*materials*), Industrials (*industrials*), Consumer Discretionary (*discretionary*), Consumer Staples (*staples*), Health Care (*health*), Financials (*financials*), Information Technology (*tech*), Communication Services (*comm*) and Utilities (*utilities*)). Variable importance is an average over all training samples. Variable importances within a model are normalized to sum to one, giving them the interpretation of relative importance for that model. As variable importances within a model can be negative, the normalized variable importance for other characteristics within that model can be greater than one. The sample consists of US CRSP stocks, excluding microcap stocks with a market capitalization smaller than the 20th percentile of stocks listed on the NYSE. The sample runs from January 1994 to December 2022.

Figure 1 — continued



one-month-ahead stock return predictions at the end of each month. Based on each model's forecasts, I sort stocks into decile portfolios using NYSE breakpoints and reassign the portfolios each month. I compute the value-weighted relative returns of holding the decile portfolios over the next month. In the end, I construct a long-short portfolio with zero net investment. The long-short portfolio buys the highest expected return stocks (decile 10) and sells the lowest (decile 1).

I evaluate three portfolio strategies to compare the profitability of the global model versus the sector models. The first sorts stocks into decile portfolios based on the global model's relative return forecasts. The second strategy performs sector-neutral portfolio sorts on the global model's return predictions. The third strategy produces decile portfolios based on the forecasts of the ten sector models.

Table 3 reports the out-of-sample performance of all value-weighted decile portfolios for the three neural network strategies. The results align with the out-of-sample predictive performance of the models on individual stock-level in Section 5.1. The global model in Panel A again dominates the sector models in Panel C in out-of-sample portfolio performance.

For the global model in Panel A, the average realized relative returns (*Avg*) generally increase monotonically throughout all decile portfolios from lowest (decile 1) to highest (decile 10). Only the difference between decile nine and the top decile is significantly larger, as the return is more than triple. The lowest decile portfolio earns a monthly value-weighted return of -0.97%, and the highest decile portfolio earns 1.73%. The long-short portfolio (*H-L*) based on the global neural network forecasts returns on average 2.71% per month (32.52% on an annualized basis). Its monthly volatility (*Std*) is 4.52% (15.66% annualized), resulting in an annualized out-of-sample Sharpe ratio (*SR*) of 2.07.

The global model with sector-neutral portfolio sorts in Panel B produces similar results but is slightly less profitable. Realized returns increase monotonically throughout all decile portfolios from lowest to highest, with a larger gap between decile nine and the highest decile. The most extreme deciles produce lower average returns than the global model with unrestricted portfolio sorts in Panel A but also experience lower monthly volatility. The lowest decile portfolio earns a monthly value-weighted return of -0.82%, and the highest decile portfolio earns 1.26%. The long-short portfolio based on the global model with sector-neutral portfolio sorts returns on average 2.08% per month (24.96% on an annualized basis). Its monthly volatility is 3.74% (12.96% annualized), resulting in an annualized out-of-sample Sharpe ratio of 1.93.

Analogous to the predictive performance of the sector neural networks out-of-sample, the portfolio strategy based on the sector models in Panel C is significantly less profitable than the global model. Realized relative returns no longer monotonically increase throughout all decile portfolios from lowest to highest. The highest decile portfolios stand out. The ninth decile should contain the stocks with the second-highest expected returns, but it generates a negative average monthly return of -0.08%. The top decile portfolio earns 0.53%, less than a third of the global model's top decile return. The lowest decile portfolio earns a monthly value-weighted return of -0.46%, half of the global model in Panel



Figure 2: Ranked variable importance across models

This figure presents the ranked variable importance for the 100 most influential characteristics across all models. Characteristics are ordered based on the sum of their ranks over all models, with the most influential characteristics on top and the least influential on bottom. Columns correspond to individual models, and color gradients within each column indicate the model's most influential (dark blue) to least influential (white) characteristics. The sample consists of US CRSP stocks, excluding microcap stocks with a market capitalization smaller than the 20th percentile of stocks listed on the NYSE. The sample runs from January 1994 to December 2022.

Table 3: Performance of neural network portfolios

This table summarizes the out-of-sample performance of decile portfolios based on three different neural network strategies. All stocks are sorted into decile portfolios based on their predicted relative returns (with the market component removed) for the next month. The table reports value-weighted returns for all decile portfolios and a zero-net-investment portfolio (*H-L*) that buys the highest expected return stocks (decile 10) and sells the lowest (decile 1). For each portfolio, the table presents the average predicted monthly returns in percentages (*Pred*), the average realized monthly returns in percentages (*Avg*), their standard deviations in percentages (*Std*), and annualized Sharpe ratios (*SR*). Panel A reports results for the global model. Panel B reports results for decile portfolios based on the global model's return predictions but with sector-neutral portfolio sorts. Panel C reports results for decile portfolios based on return predictions from the sector models. The sample consists of US CRSP stocks, excluding microcap stocks with a market capitalization smaller than the 20th percentile of stocks listed on the NYSE. The sample runs from January 1994 to December 2022.

Panel A: Global model											
	Low	2	3	4	5	6	7	8	9	High	*H–L*
Pred	-1.22	-0.57	-0.27	-0.03	0.18	0.39	0.63	0.89	1.27	2.69	3.91
Avg	-0.97	-0.44	-0.25	-0.08	-0.02	0.04	0.14	0.22	0.51	1.73	2.71
Std	2.85	2.13	2.25	2.10	2.26	2.17	2.32	2.84	3.20	3.35	4.52
SR	-1.18	-0.71	-0.39	-0.13	-0.02	0.07	0.20	0.26	0.55	1.79	2.07
Panel B: Global model with sector-neutral portfolio sorts											
	Low	2	3	4	5	6	7	8	9	High	*H–L*
Pred	-1.11	-0.53	-0.26	-0.04	0.16	0.36	0.57	0.83	1.15	2.30	3.41
Avg	-0.82	-0.50	-0.25	-0.21	0.00	0.07	0.11	0.21	0.39	1.26	2.08
Std	2.45	2.02	2.06	2.11	2.35	2.12	2.53	2.70	2.88	3.00	3.74
SR	-1.16	-0.85	-0.42	-0.34	0.01	0.11	0.15	0.26	0.47	1.45	1.93
Panel C: Sector models											
	Low	2	3	4	5	6	7	8	9	High	*H–L*
Pred	-2.85	-1.32	-0.64	-0.08	0.42	0.89	1.43	2.21	3.33	6.29	9.14
Avg	-0.46	-0.25	-0.15	-0.03	-0.04	-0.13	0.00	0.23	-0.08	0.53	0.99
Std	2.49	2.19	2.12	2.18	2.14	2.29	2.36	2.27	2.52	2.61	3.00
SR	-0.64	-0.40	-0.25	-0.04	-0.07	-0.20	0.00	0.34	-0.12	0.70	1.14

A. The long-short portfolio based on the sector neural networks returns on average 0.99% per month (11.88% on an annualized basis). Its monthly volatility is 3.00% (10.39% annualized), resulting in an annualized out-of-sample Sharpe ratio of 1.14.

Across the three strategies, the predicted out-of-sample returns (*Pred*) do not match the realized returns (*Avg*) very closely, except for the pattern of increasing returns throughout the deciles. The neural networks produce extreme return forecasts for the lowest and highest decile portfolios and overstate realized returns. This is particularly true for the sector models in Panel C. Their predicted monthly long-short (*H-L*) portfolio return based on neural network forecasts amounts to 9.14%, almost an order of magnitude larger than the realized return of 0.99%. The effect is less pronounced in the global model in Panel A. It predicted a monthly return of 3.91% for the long-short portfolio, compared to a realized return of 2.71%. The overly extreme portfolio return predictions by the sector models can be caused by overfitting. As demonstrated by the results in Section 5.1, the complex neural networks struggle with sector-specific out-of-sample predictions on small sample sizes and low observations-to-parameter ratios.

Table 4 reports further statistics on the out-of-sample performance of the three neural network strategies, focusing on the long-short portfolio returns. Panel A compares long-short portfolio returns across different out-of-sample periods. The average monthly returns over the full out-of-sample (OOS) period from January 1994 to December 2022 are the same as reported in Table 3. I further show mean returns over two subsamples: the first half of the out-of-sample period from January 1994 to December 2008 and the second half from January 2009 to December 2022. Panel A additionally reports associated t-statistics (*t-stat*).

The global model achieves statistically significant mean returns over the entire out-of-sample period and both subsamples. Long-short portfolios for the global model with unrestricted portfolio sorts outperform sector-neutral portfolio sorts over both subsamples. More importantly, the global neural network dominates the sector-specific models in all samples, both in terms of mean return and risk-return profile based on the associated t-statistics. However, its outperfor-

Table 4: Statistics of neural network long-short portfolios

This table summarizes the out-of-sample statistics of the value-weighted long-short portfolios formed from different neural network model return predictions. All stocks are sorted into decile portfolios based on their predicted relative returns (with the market component removed) for the next month. A long-short portfolio buys the highest expected return stocks (decile 10) and sells the lowest (decile 1). Results are reported for the global model, the global model with sector-neutral portfolio sorts, and the sector models. Panel A presents the average value-weighted monthly full sample mean return and average monthly sub-sample mean returns with associated t-statistics (*t-stat*). Panel B reports the average CAPM alphas and average Fama and French (2015) five-factor model (FF5) alphas, corresponding Newey and West (1987) adjusted t-statistics with six lags ($t\text{-}stat_\alpha$), and corresponding R^2. The sample consists of US CRSP stocks, excluding microcap stocks with a market capitalization smaller than the 20th percentile of stocks listed on the NYSE. The sample runs from January 1994 to December 2022.

	Global model	Global model + sector-neutral sorts	Sector models
Panel A: Percentage returns			
Mean 1994–2022	2.71	2.08	0.99
t-stat	*11.16*	*10.36*	*6.14*
Mean 1994–2008	3.24	2.66	0.70
t-stat	*9.55*	*9.08*	*3.06*
Mean 2009–2022	2.14	1.46	1.29
t-stat	*6.24*	*5.49*	*5.80*
Panel B: Risk-adjusted performance			
CAPM alpha (%)	2.57	2.00	0.98
$t\text{-}stat_\alpha$	*8.80*	*9.55*	*5.78*
R^2	*0.03*	*0.02*	*-0.003*
FF5 alpha (%)	2.44	1.90	0.85
$t\text{-}stat_\alpha$	*8.80*	*8.78*	*5.98*
R^2	*0.14*	*0.12*	*0.04*

mance is significantly more substantial in the first half of the out-of-sample period, when it earns more than four times as much as the sector neural networks.

The global model performs better from 1994 to 2008 with a 3.24% monthly long-short return and an associated t-statistic of 9.55 compared to a 2.14% return from 2009 to 2022 with an associated t-statistic of 6.24. This is analogous to the results of Blitz et al. (2023), who find a weaker out-of-sample performance for machine learning models in their later subsample after 2004. They base this result partly on the strength of the Size factor over different time periods. Size is the most important predictor for returns in my global neural network. According to Blitz et al. (2023), the stronger performance of the global model in the earlier subsample can be attributed to the excellent performance of the Size factor in earlier periods. The Size factor starts to perform worse in the last 20 years, which weakens the profitability of the global neural network in the second subsample.

The long-short portfolio based on the sorted predictions of the sector models underperforms the global model, particularly in the first half of the out-of-sample period. It achieves a monthly mean return of 0.7% with an associated t-statistic of 3.06. Although still less profitable than the global model, the sector-specific neural networks improve in the second subsample with an average return of 1.29%. Associated t-statistics are not far apart, with 5.80 for the sector models and 6.24 for the global model. Therefore, both strategies are similar in risk-return profile over the second half of the out-of-sample period.

Two factors can drive the improved performance of the sector models in the later period of the sample. First, not all sector neural networks rely on Size as an essential signal to the same extent as the global model. Therefore, their profitability is not as dependent on the performance of the Size factor. Second, in the earlier part of the sample, the training samples for the sector-specific neural networks are tiny, with a very low observations-to-parameters ratio. This makes it more challenging to estimate coefficients in a neural network without overfitting on the training data.

Table A1 in the Appendix reports the out-of-sample performance of individual GICS sector long-short portfolios based on the sorted predictions from the ten sector-specific neural networks.

Panel B of Table 4 summarizes the risk-adjusted performance of the long-short portfolios for each strategy based on factor pricing models. I report alphas on top of the Capital Asset Pricing Model (CAPM) and the Fama and French (2015) five-factor model (FF5) with associated Newey and

West (1987) adjusted t-statistics with six lags ($t\text{-}stat_\alpha$) and adjusted R^2 with respect to each factor model.

The risk-adjusted performance yields similar results as the raw long-short returns with a superior global model. All three neural network strategies achieve statistically significant alphas with t-statistics ranging from 5.78 for the sector models on the CAPM to 9.55 for the global model with sector-neutral portfolio sorts on the CAPM. The global model produces the highest alphas, with 2.57% on the CAPM and 2.44% on FF5. The sector-neutral portfolio sorts slightly lower the alphas of the global model to 2.00% on top of the CAPM and 1.90% on top of FF5. The sector-specific models span the lowest alphas, with 0.98% on the CAPM and 0.85% on FF5. The CAPM barely has any explanatory power on the average long-short returns of neural network forecasts, with R^2 never exceeding 0.03 for the global model. The five-factor model explains as much as 14% of the variation in the long-short portfolio based on the global model's forecasts. Unsurprisingly, the Size factor is the statistically most significant factor in regressions of portfolio returns on the five-factor model for all neural network strategies.

The results of Tables 3 and 4 are illustrated in Figure 3. It plots the cumulative log returns of the value-weighted long and short sides for the three neural network strategies in the out-of-sample period. The long side buys the stocks in the highest decile portfolio and the short side sells the lowest decile portfolio. Therefore, returns on the short portfolio are the relative returns of the lowest decile stocks multiplied by -1. The cumulative performance is cut off in December 2008 and restarted in January 2009 to present differences in cumulative returns between the two halves of the out-of-sample period.

The global model consistently dominates the sector models over time. However, its outperformance is mainly in the first half of the out-of-sample period and tapers off thereafter. The cumulative returns of all three strategies follow similar patterns in the second half of the out-of-sample period. The long-short spreads are still larger for the global neural network after 2008, but the magnitude of the returns relative to the sector models is smaller. Sector-neutral portfolio sorting prevents the global model from outperforming the sector neural networks in the second half of the out-of-sample period. The performance of long-short portfolios for the global model is not predominantly based on the short side, which would raise questions about practical implementation due to shorting frictions.

The return series of the global model's long portfolio is strong initially and starts to shift after 2000. It still delivers positive results, but with higher volatility, the overall magnitude of relative returns is lower. The stocks in the top decile of the global neural network's forecasts cumulate double the returns in the eight years from 1994 to 2001 than in the following seven years. The second half of the out-of-sample period from 2009 to 2023 accumulates roughly the same returns as the first eight years. Apart from the shift in the return series during the dot-com bubble crash in 2001, global shocks such as the financial crisis of 2008 and 2009 or the COVID-19 pandemic in early 2020 did not cause significant portfolio downturns.

The short side of the global model generates positive returns but underperforms the long portfolio in both subsamples. Its positive performance is mainly due to the dot-com bubble crash from 2000 to 2002. Apart from this period, cumulative returns increase only slightly or move sideways over extended periods in the plot.

The long and short sides of the portfolios for the global model with sector-neutral portfolios generally follow the same pattern. The magnitude of returns and the long-short spread are smaller. The shift in the return series after 2000 is more pronounced, and the dot-com bubble crash causes a portfolio downturn.

The top and bottom decile portfolios based on the sorted predictions of the sector neural networks do not perform well in the first half of the out-of-sample period. The long portfolio generates no significant returns after 1999 and wipes out all accumulated returns in the dot-com bubble crash. The short portfolio benefits from this crisis but otherwise does not generate any substantial returns. It is the only portfolio to accumulate negative returns in parts of the out-of-sample period. After 2008, the cumulative returns of the long and short portfolios recover but never reach the magnitude of the global model.

Appendix A2 plots the cumulative log returns of top and bottom decile portfolios sorted on the out-of-sample return forecasts of individual sector-specific neural network models.

5.4. Benchmarking with OLS sector models

Sector-specific neural networks underperform a benchmark in the form of sector-specific ordinary least squares (OLS) models in the full sample. A long-short portfolio based on the sorted predictions of OLS sector models generates a higher value-weighted relative return than the sector neural networks. However, the OLS outperformance comes only from the first half of the out-of-sample period. The trade-off between stable model estimation and capturing sector-specific complex interactions is particularly relevant for small sector samples. Forecasts from the global neural network on the pooled data across sectors remain more profitable than the OLS models.

Sector models underperform the global model regarding stock-level forecasting and long-short portfolio performance, even when sector-specific breakpoints sort the global model predictions. As a first step towards understanding the reasons for this observation, I compare the sector-specific models with another machine learning technique. The sector models used in this thesis build on a neural network architecture with three hidden layers (NN3). I compare these with sector-specific models based on a simple linear predictive regression model estimated by ordinary least squares. The methodology remains the same; only the machine learning model for predicting returns from stock-specific signals changes.

Table 5 compares the out-of-sample performance of long-short portfolios based on the sorted predictions from the two

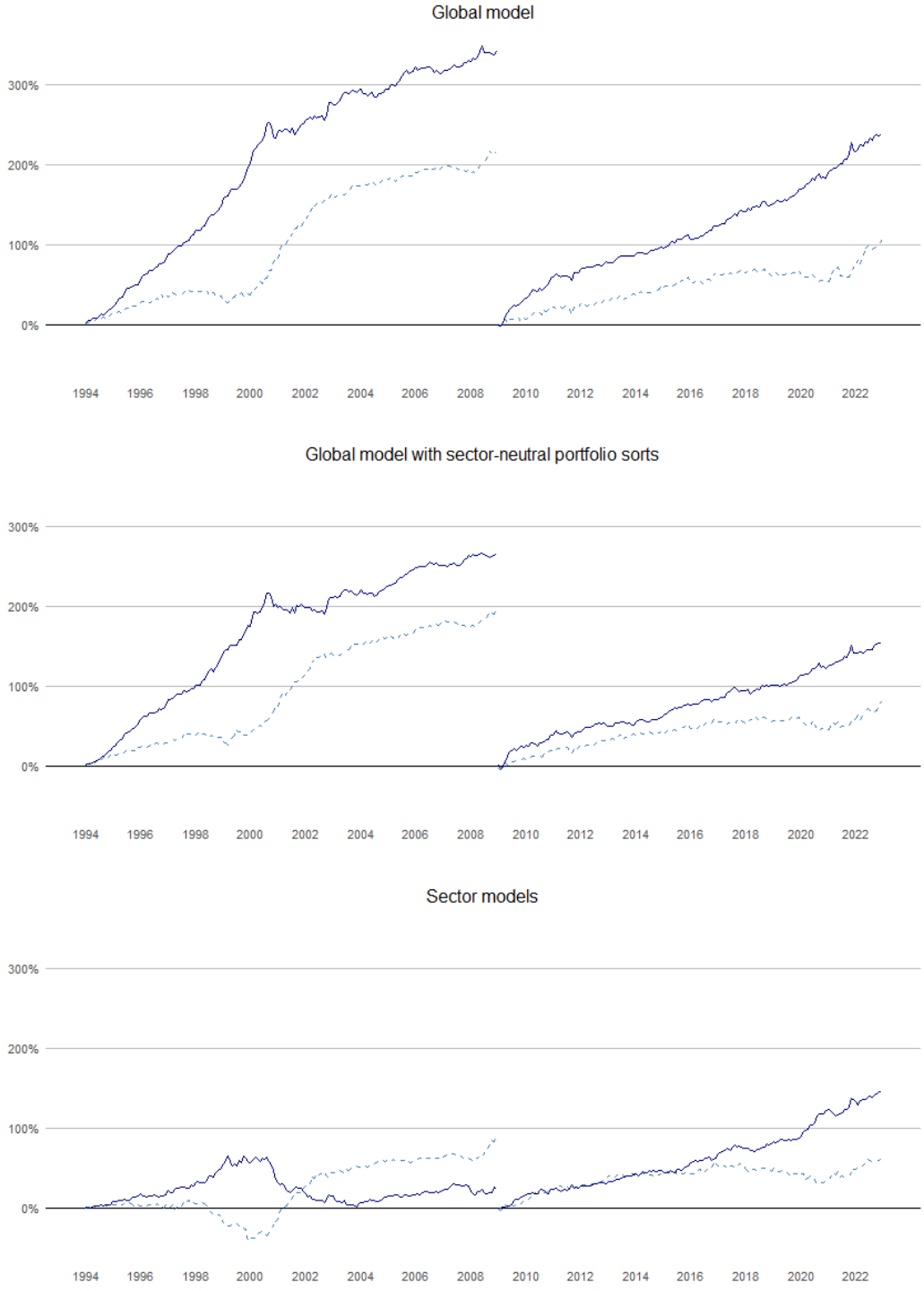


Figure 3: Cumulative performance of top and bottom decile portfolios

The figure plots the cumulative log returns of top and bottom decile portfolios sorted on the out-of-sample neural network return forecasts. Each month, stocks are sorted into value-weighted decile portfolios based on the predicted relative returns (with the market component removed) from the three neural network strategies. The solid and dash lines represent long (top decile) and short (bottom decile) positions, respectively. For the short position, the monthly relative returns of the bottom portfolio are multiplied by -1. The figure includes plots for the cumulative returns of the global model, the global model with sector-neutral portfolio sorts, and the sector models. The sample consists of US CRSP stocks, excluding microcap stocks with a market capitalization smaller than the 20th percentile of stocks listed on the NYSE. The sample runs from January 1994 to December 2022. The cumulative performance is cut off in December 2008 and restarted in January 2009 to present differences in cumulative returns between the two sub-samples.

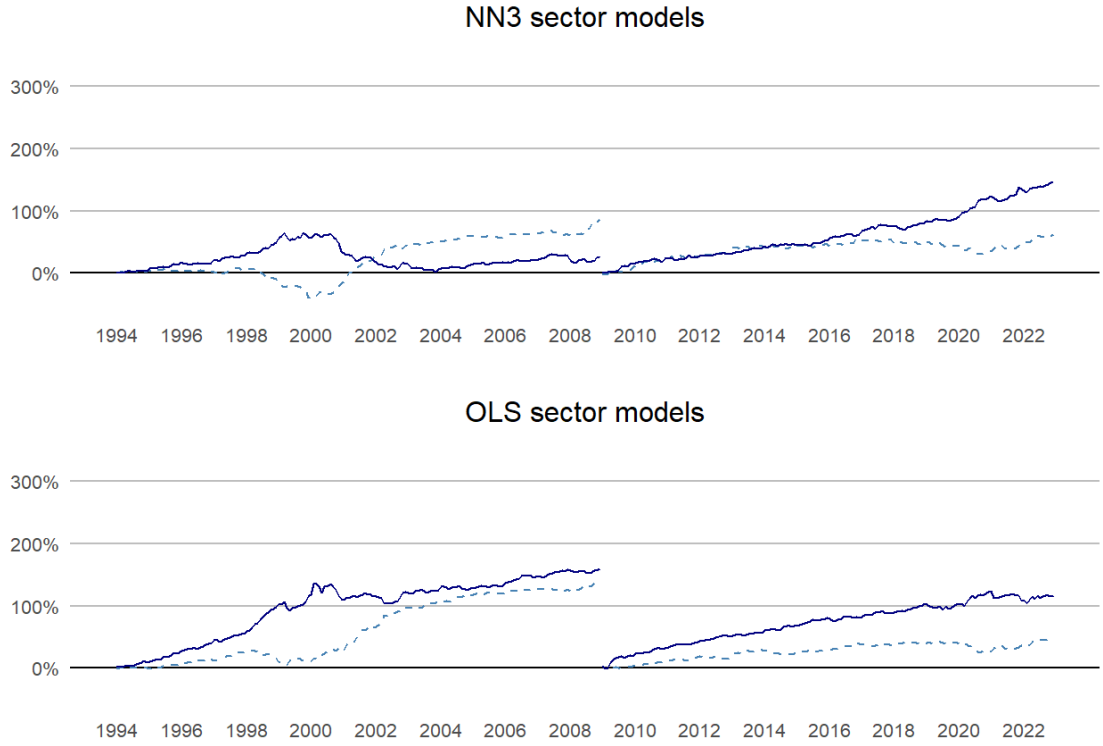
Table 5: Performance of NN3 sector models versus OLS sector models

This table compares the out-of-sample performance of the value-weighted long-short portfolios formed from different machine learning sector model return predictions. Results are reported for the NN3 sector models (same sector models as in Table 4 and Figure 3) and OLS sector models. All stocks are sorted into decile portfolios based on their predicted relative returns (with the market component removed) for the next month. A long-short portfolio buys the highest expected return stocks (decile 10) and sells the lowest (decile 1). Panel A compares the average value-weighted monthly full sample mean return and average monthly sub-sample mean returns with associated t-statistics (t-stat). Panel B visualizes the cumulative log returns of top and bottom decile portfolios sorted on the out-of-sample return forecasts of the two machine learning strategies. The solid and dash lines represent long (top decile) and short (bottom decile) positions, respectively. The sample consists of US CRSP stocks, excluding microcap stocks with a market capitalization smaller than the 20th percentile of stocks listed on the NYSE. The sample runs from January 1994 to December 2022.

Panel A: Percentage returns

Sector Models	Mean 1994-2022	*t-stat*	Mean 1994-2008	*t-stat*	Mean 2009-2022	*t-stat*
NN3	0.99	*6.14*	0.70	*3.06*	1.29	*5.80*
OLS	1.43	*7.80*	1.78	*6.41*	1.05	*4.53*

Panel B: Cumulative performance of top and bottom decile portfolios



different machine learning methods. Panel A reports average monthly value-weighted out-of-sample returns with associated t-statistics (*t-stat*). The sector neural networks (NN3) returns are the same as in Panel A of Table 4. Panel B visualizes the cumulative out-of-sample log returns of the value-weighted long and short sides for the two machine learning models.

The *OLS* sector models outperform the neural network sector models over the whole out-of-sample period from January 1994 to December 2022. They achieve average monthly long-short portfolio returns of 1.43% with an associated t-statistic of 7.80. However, the results differ for the two different subsamples. The outperformance of the OLS models is based solely on the first half of the out-of-sample period from 1994 to 2008. The OLS sector models generate significantly higher long-short returns than the NN3 sector models during this period. The monthly relative returns of 1.78% for the OLS models are more than double the 0.70% achieved by the sector models. The picture changes when looking at the later subsample from 2009 to 2022. In this period, OLS models underperform NN3 models, with average monthly returns of 1.05% and an associated t-statistic of 4.53. Panel B illustrates these results. The cumulative log returns for the OLS long and short portfolios remain positive throughout the out-of-sample period. They perform more reliably in the early years, with only a tiny long-short spread. As with all other machine

learning models, the gains for the short portfolio are mainly due to the dotcom bubble crash. The biggest difference between the OLS and NN3 sector models is the long portfolio in the first half of the out-of-sample period. The cumulative returns of the top decile stocks based on OLS forecasts are about an order of magnitude higher than their NN3 counterparts. In the second subsample from 2009 to 2022, the returns of the OLS long and short portfolios follow a similar pattern to the NN3 returns, with a slight underperformance.

These results demonstrate a trade-off between robust model estimation and complex model architecture to capture sector-specific nonlinearities and variable interactions. A simple linear regression model such as OLS requires only a single parameter for each of the 212 signals. The parameters for the optimization problem are derived from a closed-form solution. Therefore, OLS models benefit from a higher ratio of observations to parameters and more stable model parameters. Their obvious disadvantage is the inability to capture non-linearities and variable interactions. The number of estimated coefficients in each of my neural network models is 7,489. These numbers result in a low observations-to-parameter ratio on small data sets, such as the individual sector samples, especially at the beginning of my sampling period when the training samples are the smallest. In low signal-to-noise problems like stock return prediction, complex machine learning models such as neural networks tend to overfit noise rather than extract signals. The NN3 sector models cannot exploit their advantage of being able to capture complex sector-specific relationships between signals and future returns. This can lead to the poor predictive performance of neural networks for small sectors and low portfolio returns in the first half of the out-of-sample period. The pooling of data across sectors for the global model improves the ratio of observations to parameters. This explains why it does not suffer from the same problems as the sectoral models. The global neural network is still able to capture nonlinearities and variable interactions across sectors and thus significantly outperforms OLS models in the full sample, in line with previous research (e.g., Azevedo and Hoegner (2023), Blitz et al. (2023), and Gu et al. (2020)).

Table A3 in the Appendix summarizes the out-of-sample statistics of the value-weighted long-short portfolios formed from different sector-specific OLS model return predictions.

5.5. Sector allocation as a return driver

The global neural network demonstrates some out-of-sample sector allocation power. In the cross-section of sectors, it correctly predicts higher relative returns for the most profitable sectors and lower returns for the least profitable sectors. This allows the global model to generate higher returns in the long (top decile) portfolio when portfolio sorting is not sector-neutral.

As seen in Table 4, sector-neutral portfolio sorts worsen the profitability of long-short portfolios sorted on the out-of-sample return forecasts of the global neural network. To better understand this difference in performance, I briefly evaluate the global neural network's potential out-of-sample sector allocation power. Table 6 summarizes my results. For each of the ten GICS sectors, Panel A reports the average value-weighted realized monthly relative return compared to the average predicted return from the global model. In addition, I report the average monthly allocation to the respective sector in the top decile portfolio sorted based on forecasts from the global model. I focus on the long portfolio, the stronger driver of the global model's profitability than the short portfolio. Panel B plots the cumulative value-weighted relative log returns per sector over the out-of-sample period from January 1994 to December 2022.

Information Technology (*tech*) is the best-performing sector over the entire out-of-sample period, with a monthly relative return of 0.23%. Health Care (*health*) follows in second place with a return of 0.09%, and Energy (*energy*) is in third place with a return of 0.01%, thanks to solid gains after 2020. All other sectors generate negative value-weighted relative returns out-of-sample.[5] The worst-performing sectors are Utilities (*utilities*), with an average monthly return of -0.24%, and Communication Services (*comm*), with a return of -0.35%. Apart from Information Technology, the global model's return predictions are not very close to the realized returns, and they overstate expected returns. Still, it performs relatively well in cross-sectionally classifying the sectors into the correct extremes. The global model correctly identifies the two best-performing sectors and predicts Utilities and Communication Services to be among the three worst-performing sectors.

As a result, the global model shows some (limited) out-of-sample sector allocation power. The long portfolio based on global neural network forecasts has high average allocations to Information Technology (22.93%) and Health Care (15.09%). It has an average allocation of less than 5% to each of the three worst-performing sectors: Materials, Utilities, and Communication Services. The global model generates higher returns than sector-neutral portfolios by overweighting more profitable sectors and underweighting less profitable sectors in the top decile portfolio.

6. Conclusion

I examine the difference in predictive power for the cross-section of US stock returns between a global machine learning model and sector-specific models. Based on their strong performance in previous research, I use neural networks as the machine learning models. The global neural network is trained on the full sample of stocks, while the sector neural networks are trained on ten different GICS sectors. The global model consistently outperforms the sector models out-of-sample in terms of predictive accuracy and profitability. It

[5] Equally-weighted results (not reported) show that almost all sectors have positive relative returns. On average, stocks with smaller market capitalization generate higher relative returns. This further demonstrates the strong performance of the Size factor in the sample and justifies the high variable importance of Size in the global neural network.

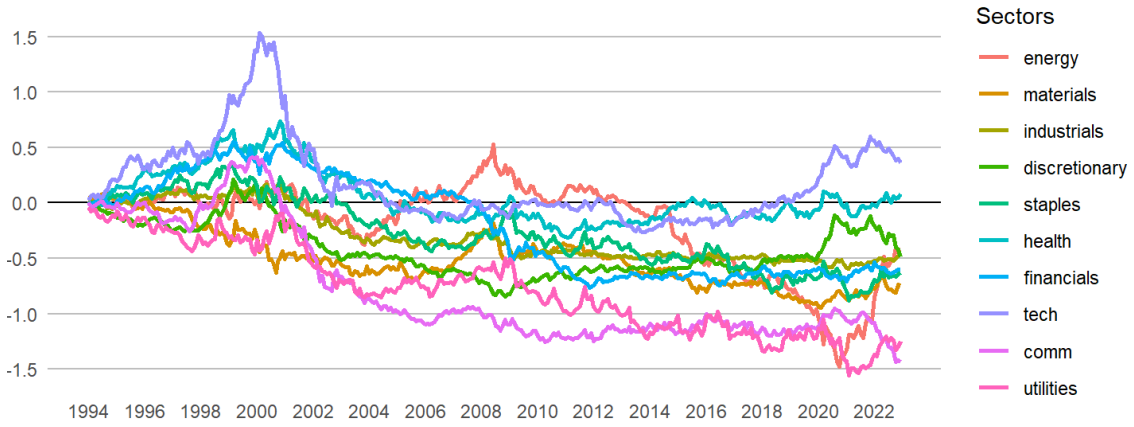
Table 6: Sector allocation power of global neural network model

This table summarizes parts of the out-of-sample sector allocation power of the global neural network model. Panel A compares the average monthly realized relative returns (with the market component removed) per sector to the average monthly predicted relative returns from the global model over the out-of-sample period. The sectors in Panel A are ranked in descending order of realized returns. Panel A additionally reports the average monthly allocation to each sector of the long (top decile) portfolio sorted based on the global model's return predictions for the next month. Panel B plots the cumulative relative log returns per sector over the out-of-sample period. All returns are value-weighted. The out-of-sample period runs from January 1994 to December 2022. The sample consists of US CRSP stocks, excluding microcap stocks with a market capitalization smaller than the 20th percentile of stocks listed on the NYSE.

Panel A: Percentage returns and long portfolio allocation

Sector	Realized returns	Predicted returns	Long portfolio allocation
Information Technology	0.23	0.26	22.93
Health Care	0.09	0.17	15.09
Energy	0.01	0.06	7.21
Consumer Discretionary	-0.10	0.14	14.52
Consumer Staples	-0.11	-0.03	3.74
Industrials	-0.11	0.12	11.54
Financials	-0.13	0.14	12.96
Materials	-0.16	0.11	4.50
Utilities	-0.24	-0.02	3.96
Communication Services	-0.35	0.05	4.36

Panel B: Cumulative relative log returns per sector



derives most of its predictive power from Size as an input signal. A long-short portfolio based on the sorted predictions of the global model generates significant returns and Sharpe ratios over the entire out-of-sample period. The sector models generate negative out-of-sample R^2_{OOS} and their long-short portfolio returns are lower, especially in the early out-of-sample period. Complex models such as non-linear neural networks struggle to exploit their advantages on small sector-specific samples and underperform simple OLS models. The results for the global model support the recent literature on the strong predictive power of neural networks for the cross-section of stock returns.

References

Amihud, Y. (2002). Illiquidity and stock returns: cross-section and time-series effects. *Journal of Financial Markets*, *5*(1), 31–56.

Avramov, D., Cheng, S., & Metzker, L. (2023). Machine Learning vs. Economic Restrictions: Evidence from Stock Return Predictability. *Management Science*, *69*(5), 2587–2619.

Azevedo, V., & Hoegner, C. (2023). Enhancing stock market anomalies with machine learning. *Review of Quantitative Finance and Accounting*, *60*(1), 195–230.

Azevedo, V., Hoegner, C., & Velikov, M. (2024). The Expected Returns on Machine-Learning Strategies. *SSRN Electronic Journal*.

Azevedo, V., Kaiser, G. S., & Mueller, S. (2023). Stock market anomalies and machine learning across the globe. *Journal of Asset Management*, *24*(5), 419–441.

Blitz, D., Hanauer, M. X., Hoogteijling, T., & Howard, C. (2023). The Term Structure of Machine Learning Alpha. *SSRN Electronic Journal*, 1–40.

Cakici, N., Fieberg, C., Metko, D., & Zaremba, A. (2023). Machine learning goes global: Cross-sectional return predictability in international stock markets. *Journal of Economic Dynamics and Control*, *155*, 1–32.

Cavaglia, S. M. F. G., Sefton, J., Scowcroft, A., & Smith, B. (2006). Global Style Investing. *The Journal of Portfolio Management*, *32*(4), 10–22.

Chen, A. Y., & Zimmermann, T. (2022). Open Source Cross-Sectional Asset Pricing. *Critical Finance Review*, *27*(2), 207–264.

Chen, L., Pelger, M., & Zhu, J. (2024). Deep Learning in Asset Pricing. *Management Science*, *70*(2), 714–750.

Ciner, C. (2019). Do industry returns predict the stock market? A reprise using the random forest. *Quarterly Review of Economics and Finance*, *72*, 152–158.

Drobetz, W., & Otto, T. (2021). Empirical asset pricing via machine learning: evidence from the European stock market. *Journal of Asset Management*, *22*(7), 507–538.

Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, *116*(1), 1–22.

French, K. R. (2024). Current Research Returns. Retrieved November 1, 2024, from http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

Freyberger, J., Neuhierl, A., & Weber, M. (2020). Dissecting Characteristics Nonparametrically. *The Review of Financial Studies*, *33*(5), 2326–2377.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies*, *33*(5), 2223–2273.

Hanauer, M. X., & Kalsbach, T. (2023). Machine learning and the cross-section of emerging market stock returns. *Emerging Markets Review*, *55*, 101022.

Hou, K., Xue, C., & Zhang, L. (2020). Replicating Anomalies. *The Review of Financial Studies*, *33*(5), 2019–2133.

Kim, K. J., Lee, C., & Tiras, S. L. (2013). The effects of adjusting the residual income model for industry and firm-specific factors when predicting future abnormal returns. *Asia-Pacific Journal of Financial Studies*, *42*(3), 373–402.

Leippold, M., Wang, Q., & Zhou, W. (2022). Machine learning in the Chinese stock market. *Journal of Financial Economics*, *145*(2), 64–82.

Lewellen, J. (2015). The Cross-section of Expected Stock Returns. *Critical Finance Review*, *4*(1), 1–44.

Liu, X., Pong, E. S. Y., Shackleton, M. B., & Zhang, Y. (2014). Option-Implied Volatilities and Stock Returns: Evidence from Industry-Neutral Portfolios. *The Journal of Portfolio Management*, *41*(1), 65–77.

Moritz, B., & Zimmermann, T. (2016). Tree-Based Conditional Portfolio Sorts: The Relation between Past and Future Stock Returns. *SSRN Electronic Journal*, 1–81.

MSCI Inc. (2024). The Global Industry Classification Standard (GICS®). Retrieved November 1, 2024, from https://www.msci.com/our-solutions/indexes/gics

Nagel, S. (2005). Short sales, institutional investors and the cross-section of stock returns. *Journal of Financial Economics*, *78*(2), 277–309.

Newey, W. K., & West, K. D. (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *The Econometric Society*, *55*(3), 703–708.

Rapach, D. E., Strauss, J. K., Tu, J., & Zhou, G. (2019). Industry Return Predictability: A Machine Learning Approach. *The Journal of Financial Data Science*, *1*(3), 9–28.

Rasekhschaffe, K. C., & Jones, R. C. (2019). Machine Learning for Stock Selection. *Financial Analysts Journal*, *75*(3), 70–88.

Tobek, O., & Hronec, M. (2021). Does it pay to follow anomalies research? Machine learning approach with international evidence. *Journal of Financial Markets*, *56*.

U.S. Department of Labor. (2025). Standard Industrial Classification (SIC) Manual. Retrieved November 1, 2024, from https://www.osha.gov/data/sic-manual

Junior Management Science 10(3) (2025) 582-608



Junior Management Science

www.jums.academy
ISSN: 2942-1861



Beware of Bullshit – A Qualitative Study on Young Adults' Sustainability Awareness of Online Services

Robin Roskosch

University of Duisburg-Essen

Abstract

Since the surge of working in the 'home office' during the COVID-19 pandemic, information and communication technology has been seen as a problem solver for the global climate disruption, because it promises savings in emissions through digitalization. However, this is criticized for being a 'double-edged sword', as more use of ICT also means an increase in electricity consumption. This thesis poses the question if consumers using services fostering said digitalization are aware of this implication and can contribute to sustainable development by choosing wisely. Thus, this paper leads exploratory research on sustainability awareness of online services, from the perspective of both sustainably aware and technologically skilled individuals. This will be achieved by collecting qualitative data from interviews with two groups of university students from biology and computer science or information systems. The data will be evaluated using the approach of qualitative content analysis and will be investigated further with the help of the concept of bullshitting, to understand how consumers assess ambiguous and misleading sustainability claims about online services. One implication from the interviews is that both companies and consumers should strive to promote reliable knowledge on this topic, as there is a deficit in sustainability awareness of online services.

Keywords: bullshit; green IT; greenwashing; online services; sustainability

1. Introduction

Advances in information and communication technologies (ICT), as well as sustainable development, are two topics that are being actively discussed in today's society, politics, and economy. Through digitization, digitalization, and digital transformation, ICT is regularly presented as one of the possible solutions to the challenges of sustainable development. For instance, moving from analog to digital data formats with the help of *digitization* and supporting or automating processes through *digitalization*, omits filling out a transfer slip and driving to the next bank, leading to paper savings and reduced travel (Vrana & Singh, 2021, p. 4). Furthermore, Internet of Things (IoT) as an example for *digital transformation* technologies, can be used to change whole business models to optimize energy consumption (Ericsson, 2020, p. 9). Therefore ICT can be seen as an enabler, but its role as part of the problem should not be neglected (Ericsson, 2020, p. 9; Hofmann et al., 2021, p. 7). ICT can hurt environmental sustainability due to massive energy consumption as a result of increased use, toxic e-waste as a result of poor recycling, or general high demand for rare resources used in manufacturing.

A major example of how ICT can be a 'double-edged sword' is the recent rise in the implementation of home-

I would like to thank my tutor, Dr. Helge Alsdorf, for encouraging me to explore the topic of Bullshitting and Green IT in my bachelor thesis, building on the foundation of my seminar paper. His suggestions and support were greatly appreciated throughout this process. A special thanks goes to Felicitas Comfai for her ongoing motivation, valuable feedback, and careful proofreading, all of which were a great help during the writing of this thesis. I would also like to thank Ole Gulz for his proofreading and feedback, which was a great help in refining my work. I also wish to thank all the interview partners who generously shared their time and perspectives, contributing meaningful insights to this work. Lastly, I am grateful to the SITM chair for providing a platform to investigate unique and unconventional topics such as the concept of bullshitting.

DOI: https://doi.org/10.5282/jums/v10i3pp582-608

© The Author(s) 2025. Published by Junior Management Science.
This is an Open Access article distributed under the terms of the CC-BY-4.0 (Attribution 4.0 International). Open Access funding provided by ZBW.

based telework, better known as 'home office'. The ongoing COVID-19 pandemic has pushed companies to allow their employees to work from the safety of their homes, for example by hosting meetings online rather than in person, which has virtually established a further digital transformation by force. Working from home decreased the need for commuting and business travel and started the trend of a "considerable substitution of physical (professional) traffic by data traffic (digitalization)" (Kutani et al., 2021, pp. 21–22).

A study commissioned by Greenpeace estimates that 5.4 million tons of CO_2 emissions from transport could be cut per year if 40% of the employees had two home office days per week. This would be equivalent to 4% of total passenger transport emissions (Büttner & Breitkreuz, 2020, p. 14). However, as more and more people actually start to work from home and use their devices to communicate with colleagues and customers more frequently, data traffic is also increasing, which in turn requires more energy. This rebound effect is included in the estimate, but it is crucial to point out that it needs to be taken into account in the discussion of the potential for emission savings of ICT in general (Büttner & Breitkreuz, 2020, p. 5). Home office was by far not the only reason global internet traffic increased by about 40% in 2020 (Sandvine, 2020, p. 5). An increase in "video streaming, video conferencing, online gaming, and social networking" (Kamiya, 2021), as well as many other services contributed to that jump in numbers. The growth normalized to 29% in 2021 after the surge caused by the COVID-19 pandemic, but it is still almost a threefold increase in bandwidth since 2017 (TeleGeography, 2021, p. 1).

Since not everyone is an IT expert, many people seem to take it for granted that their preferred IT device (e.g. smartphone, tablet), as well as their favorite app, just work, without realizing how massive the IT infrastructure is that powers the services. They are unconsciously oblivious to the fact that one push of a button sets off 'lots of moving parts' on servers in the backend of an online service, requiring a lot more energy than their personal device does. The immense growth in usage and transmission rates accumulates to an energy demand of 1% and 1.1 – 1.4% of the total global electricity use in 2020 for data centers and data transmission networks respectively, which are magnitudes not comprehensible by the average consumer (Coroamă, 2021; ITU, 2020; Kamiya, 2021; Malmodin & Lundén, 2018; Masanet et al., 2020). This barrier to understanding could pose a threat to sustainable development driven and supported by consumers because the time for change is long overdue to mitigate the consequences of the climate disruption (Harvey, 2020).

Consumer practices, considered pro-environmental or Green-IT, that have been recommended so far include: double-sided printing, avoiding printing by using digital formats, turning off devices or features when not in use, and recycling or reusing old devices (Leung et al., 2018, p. 5; Molla et al., 2014, p. 133; Nash and Wakefield, 2021, p. 1). However, they all share a limitation, which becomes very apparent in a time when online services are omnipresent. All of these practices only allow users to take effect in their immediate sphere of influence, basically on or with their own devices. This means, that services that one does not maintain themself are out of reach. The ICT provider (Ericsson, 2020, p. 15) recommends that consumers make it clear to companies that they care about their carbon footprint and try to use online services that comply with science-based targets (SBTs) rather than net-zero claims. They state that SBTs are to be preferred as they are in line with the Paris Agreement's goals and do not allow to count carbon offsets toward SBTs (SBTi, 2020, p. 7; SBTi, 2021, p. 8). Net-zero targets however allow carbon offsetting to achieve 'zero' emissions. Buying carbon offsets or credits sparks a lot of controversy, especially because of its reputation as a "get out of jail free card" (Gilks & Watson, 2019). Renewable energy certificates (RECs) are not the same as carbon offsets but have a similar shady reputation amongst environmental activists. Especially unbundled RECs are criticized, because with those a company does not need to source the energy from the same grid as the origin specified in the certificate (Cook & Jardim, 2019; U.S. EPA, 2021). Which, with a bit of common sense, do not sound like directly contributing to sustainability.

Without knowing the difference between the aforementioned terms, sounding somewhat ambiguous, it can probably be difficult to follow the recommendations proposed by Ericsson. Not even "half of executives (43%) [...] are aware of their organization's IT [carbon] footprint" (Capgemini (Ed.), 2021, p. 10). This is likely attributable to the circumstance that greenhouse gas (GHG) assessment of the resources needed to run an online service are rather scarce, because the "IT-Services and underlying ICT infrastructures are complex constructs" (Grimm et al., 2013, p. 10). Knowing that, it might be even more difficult for consumers to make an educated environmentally aware decision before using online services. The point could be made that they run the risk of falling for the 'bullshit' produced by companies in the ICT industry.

The term *bullshit* is commonly known as a swear word and vulgar language but was introduced into the scientific discussion by philosopher Harry G. Frankfurt, who defined a concept behind the term and thus established the research on bullshit. The act of bullshitting can be described as a communication practice in which information is made up to hide one's ignorance. To Frankfurt (2005, pp. 33–34), the most salient feature of bullshitting is the "lack of connection to a concern with truth" (pp. 33 – 34). Frankfurt's idea of bullshit, however, is not the only one. The fellow philosopher Gerald A. Cohen (2002), describes bullshitting from a different point of view, concluding its essence to be 'the unclarifiable unclarity' – meaning utterances that are vague and unclear, so that in practice their truthfulness cannot be made out immediately. To Cohen (2002, p. 8), people that are exposed to bullshit are "mere victims", not being able to spot the truth. In a similar way, as people cannot verify the actual resource consumption and emissions of online services, regardless of whether the operators are transparent about it or not.

There is crucial evidence in the scientific literature that the concept of bullshitting is useful and demonstrates its legitimacy. It has been proven to be a "legitimate psychological phenomenon" (Pennycook et al., 2015, p. 566), a "managerial communication genre" (Christensen et al., 2019, p. 1594), and a helpful method to uncover deceptive and inconsistent claims (Stevenson, 2021, p. 90). Therefore, the concept of bullshitting promises to be a fruitful method to break down the issues surrounding the sustainability awareness of online services.

Research on sustainability awareness in ICT has so far only dealt with hardware and software in the users' sphere of influence (Molla et al., 2014; Nash & Wakefield, 2021) or has not specifically defined what is understood as an IT product for end-users (Kurkoon et al., 2016). In addition, other papers only looked at products and services where a sustainability effect is advertised as a key feature, such as time-switched smart home sockets (Koo et al., 2015), which obviates the question of awareness of 'hidden' sustainability concerns. In general, most of the aforementioned papers have collected their data quantitatively among several types of groups but failed by having a one-sided approach. Most of the people surveyed in the studies are professionally or educationally involved in the field of ICT, which does not represent the average consumer who has no knowledge about the technical details of ICT (Hernandez, 2019, p. 8; Molla et al., 2014; Nash and Wakefield, 2021, p. 2). Unfortunately, studies that asked consumers did not focus on the online service part of ICT (Kern et al., 2011; Kurkoon et al., 2016) or only focused on a specific type of service (Grimm et al., 2013).

Thus, this thesis aims to address the knowledge gap regarding awareness of 'invisible' sustainability concerns of online service use, from the perspective of both a sustainably aware and a technologically skilled individual. The aim is then to investigate said awareness with the help of the concept of bullshitting and lay the groundwork to understand how consumers assess offerings convoluted by bullshit claims. In order to achieve this research goal, the following leading research question and its sub-aspects need to be answered and discussed:

> **RQ:** *How do sustainably awareness and IT knowledge affect young adults' ability to spot bullshit claims made about the carbon footprint caused by online service usage?*

Since there is no sufficiently extensive body of research available yet on this particular topic, this work needs to build a foundation and "generate [...] initial ideas" (Bhattacherjee, 2012, p. 5), as is the case with exploratory research. The basis for answering the research question consists of qualitative data provided by interviews conducted with people from two specific groups. The groups consist of German university students, where one part of the interviewees has a background in computer science or information systems, representing individuals with IT knowledge, whereas the other part comes from the field of biology. Biology students have been chosen as the 'opposite' group because they are usually expected to have more knowledge about environmental awareness and sustainability than the average student (Arshad et al., 2020; Robinson & Crowther, 2001). In the context of the interviews, IT knowledge refers specifically to the basic understanding of the infrastructure needed to provide and use an online service. In order to limit the scope, this thesis only explores sustainability awareness in relation to so-called *digital* or *online services* for consumers. Services that fall under this umbrella term offer one or more of the following features: music and video streaming, messaging, social media, video conferencing, search engines, cloud storage and online games. The important point is that, for those services to function, most of the computation needs to be done by the service provider and not by the consumer locally. This is unavoidable, especially when content is provided to the consumer. Since services targeted at consumers are of the main interest for this thesis, IT services within a business context or hardware in general will not be considered. To include the business context, it would require interviewing people who use such services at work and add another set of questions, which would not necessarily help answering the specific research question but could be part of future research. Personal hardware does not have the problem that its operation is not controlled by the consumer. With hardware, production and recycling play a greater role, but the examination of this problem is also not within the scope of the research question. Still, its energy consumption should be kept in mind, because those devices are needed to access the online services in question.

2. Conceptual Foundation

2.1. Green IT and Green IS

When discussing the contributions and negative effects of ICT on the environment, one always stumbles across the terms *Sustainable* or *Green IT* and *IS*. Which can be either understood as the research field of sustainability in ICT, or its outcome of practices developed to minimize the environmental impact of ICT. However, sustainability can be achieved in many ways and is not only limited to ICT. Therefore, it is necessary to specify how the term sustainability is defined. In research, the concept of sustainability is frequently divided into three distinct main areas or 'pillars': *social, economic* and *environmental sustainability* (Purvis et al., 2019). Social sustainability is mainly concerned with human well-being and equality, whereas the focus of environmental sustainability is on the conservation and protection of the ecosystem (Garren & Brinkmann, 2018, p. 4). The goal of economic sustainability is to facilitate economic growth and development in accordance to the other pillars (Purvis et al., 2019, pp. 686–689). Despite the focus on different problem areas, all pillars share the core idea of sustainability, which dates back to the 17th and 18th centuries (Purvis et al., 2019, p. 682; Warde, 2011). Most notably the term was coined by the forestry pioneer Carlowitz from around that time (Warde, 2011). He described

the approach of cutting and planting enough wood to allow it to grow back in time as sustainable (Warde, 2011, p. 162). In theory sustainability then means: not taking more than is feasible, reaching a state in a system where consumption, regeneration and disposal are at an equilibrium (Ben-Eli, 2018, p. 1339).

The word 'green' in Green IT/IS aims at environmentally friendly technologies and practices, so it is primarily in line with the idea of environmental sustainability (Loeser, 2013, p. 5). Overall 'green' has been the more popular prefix than 'sustainable' in the past, but generelly they refer to the same idea. Therefore the Green IT/IS terminology will be adapted for this thesis (Loeser, 2013, p. 3). In its infancy Green IT/IS was not a precisely defined concept (Molla et al., 2009), but with a steadily growing body of research many endeavors have been undertaken to determine the important parts of the concept. As a means to summarize all this knowledge, Loeser (2013) has made it his task to combine all the important core aspects of the works published to date into one definition.

One of the main findings of this concretization is that the two terms Green IS and Green IT usually describe different partial aspects and must therefore be defined separately (Loeser, 2013). As a comprehensive concept, IS describes "the technology components and human activities related to the management and employment process of technology within the organization" (Chen et al., 2010, p. 237), whereas IT is understood as the technological fundament for IS which provides the necessary hard- and software artifacts (Ijab et al., 2010, p. 434). Following the example of the distinction between IT and IS, Green IT can be seen as a subfield of the more holistic concept of Green IS. The "cross-functional characteristics of Green IS" (p. 6) are especially emphasized by Loeser (2013), as it combines issues of technology and organization. The differences of Green IT/IS are also expressed in their range of effects, especially in terms of management. While Green IT practices are rather utilized at the operational and tactical level of IT departments, Green IS deals with far-reaching and long-term decisions at the strategic level that influence the direction of the business and the organization as such (Ijab et al., 2010, p. 441; Loeser, 2013, p. 6).

On a high level Green IT is a label for technologies and processes that have a reduced negative environmental impact compared to traditional ones, in short: they are environmentally friendly (Loeser, 2013, p. 5). This understanding can be manifested in three stages according to Loeser (2013, p. 5): First, environmental criteria should be taken into account when acquiring new IT equipment and services. Second, the operation of IT infrastructure in data centers and offices should be energy efficient. And finally, IT equipment passing the end of its life cycle should be disposed correctly or recycled. These rather operational practices can be categorized as the *1st degree environmental impact of IS (Loeser, 2013, p. 6)*. With Green IS, sustainability and environmental friendliness can be achieved through business process reengineering and environmental management systems. This is referred to as the *2nd degree*. Furthermore, the *3rd degree* is concerned with the innovation of end-user products and services, as well as tracking their emissions and resource demands.

1st and 3rd are of specific interest for this thesis, as the former is a black box to the end-user, because it involves making already established online services 'greener' in their operation. While the latter is directed toward the end-user by offering innovative services, claiming to contribute in an environmentally sustainable way. To reduce the 1st degree impact, big players of their respective service niche, like Google, Facebook, Netflix and Spotify, either source their energy from renewables, invest directly in renewable energy power plants, fund sustainability projects or offset emissions with RECs or carbon offsets (Facebook, 2020; Google, 2021a; Netflix, 2021; Spotify, 2021). Products and services reducing the 3rd degree impact for end consumers are for instance any smart home devices that promote to reduce energy consumption (Koo et al., 2015, p. 68; Loeser, 2013, p. 8). Furthermore, services can be enhanced with functions that should foster sustainable decisions, e.g. Google listing the carbon footprint next to the search results when searching for a flight (Google, 2021b). Another example can be services that completely market themselves as directly contributing to environmental sustainability, like the search engine Ecosia, which advertises to plant a tree for every search query on their site (Ecosia, 2022).

2.2. Awareness of Green IT and Green IS

The whole concept of Green IT/IS, with its strategic, tactical, and operational orientation, as well as its three degrees of impact feels very grand. But looking past the challenges regarding its implementation in organizations, it also has its influence on the level of an individual. To get an idea about research on the individual's awareness of Green IT/IS a brief literature review has been conducted. It has been noted by Dalvi-Esfahani et al. (2020, p. 2) that most of the earlier research was specific to an organizational context. Since the focus of this work is on the individual in the role of the end-user, papers with an organizational context were not specifically studied.

Kurkoon et al. (2018) conducted a survey among Thai people familiar with technology products regarding their intention to purchase Green IT products. Those products had aspects about "energy efficiency, reduced greenhouse gas emission, recycling and other eco-beneficial characteristics" (Kurkoon et al., 2018, p. 1). After evaluating 618 responses, they concluded that environmental concern is present, but knowledge about Green IT products is poor. However, they had no clear differentiation between a Green IT product and a service, despite using both terms. Hernandez (2019) investigated objective and perceived knowledge of Green IT among 288 Philippine computing students using a questionnaire. He assessed them to have average Green IT knowledge (e.g., reusing old devices, using power management, reduce paper and travelling by using digital alternatives, etc.) but are not very well informed about the

implication of resource use and energy efficiency of compute device in their development and production. Knowledge regarding online services was not part of the observation. Dalvi-Esfahani et al. (2020) looked at the individual Green IT practice of Malaysian students, who already had experience with Green IT. The result of analyzing 262 responses was that the intent to practice Green IT was found to be driven by attitude, perceived behavioral control, and personal norm. A link to online services was also not present in this study. Leung et al. (2018) surveyed students from multiple countries, regarding their Green IT knowledge and practice. In addition, unlike the studies mentioned above, question about music and video streaming services were also a part of the questionnaires. Analysis of the gathered data indicates that knowledge about sustainable ICT practices influences the individual confidence in the ability and controllability of a particular behavior, also referred to as personal behavior control, meaning an increase of Green IT information increases the perception that Green IT practice is not difficult and practicing Green IT is within reach.

Molla et al. (2014) and Nash and Wakefield (2021) investigated the Green IT beliefs and practices among Australian and mostly North American IT professionals respectively. The evaluation of 322 survey responses by Molla et al. (2014) showed that IT professional believe that IT can resolve environmental problems, but only commit to simple Green IT practices. After collecting results from 157 surveys, Nash and Wakefield (2021) further discovered that self-identity in a way motivates the intention to practice Green IT, to the extent that it is part of the identity.

Koo et al. (2015) focused their research on South Korean consumers perceived usefulness of smart green IT devices that are supposed to help reduce electricity consumption. After evaluating 100 questionnaire responses, they concluded that motivational values may influence perceived usefulness, in turn positively affecting consumers' green IT usage behavior. Nevertheless, the study only narrowly looked at people already using that specific device, leaving out other products and services. Kern et al. (2011) also directed their research at consumers by performing an experiment to check whether the energy consumption of software can be improved through user configuration. They found a measurable difference, but if consumers would translate this knowledge from theory into practice was left up to debate.

From the collected literature, it can be inferred that the main concern of most studies was the examination of the individuals' Green IT adoption behavior. The lesser of those studies focused on the context outside of organizations or people who have no connection to the field of IT, professionally or in studies. Moreover, the focus was rather on practicing Green IT as an individual in the direct sense, rather than practicing it through assessing the services by their environmental friendliness or 'hidden' sustainability concerns as said in the introduction. In that sense, the individual Green IT awareness was not discussed in the context of companies and their services that claim to be green. Interestingly, Kern et al. (2011, p. 208) assumed that users could contribute to a sustainable internet by gathering information about the energy consumption of accessed servers, but it is not specified how and where they get the information from.

2.3. Online Service

In scientific literature, the word 'online service' is used in many different contexts and varies slightly in the characteristics of its definition and what it actually describes. This chapter provides an overview of commonly used terms, their meaning, and finally a unified definition for use in this thesis. Terms that are often used together or synonymously are: web service, e-service, digital service, and of course online service.

The term 'web service' emerged from computer science, as a designation in software development for functions invokable over the Internet (Baida et al., 2004, p. 2). Most often, a web service can be a technical vehicle to provision an e-service (Baida et al., 2004, p. 2; Chaudhri, 2003, p. 70; Hull et al., 2003, p. 2). According to the Council of the European Union (2011, p. 5) an e-service or 'electronically supplied services' are "impossible to ensure in the absence of information technology" (p. 5) and "involve[s] minimal human intervention" (p. 5), which is due to the fact that they are only accessible via the Internet or other networks and usually have a high degree of automation. Kvasnicova et al. (2016, p. 193) further define aspects such as the non-physical nature of an e-service, the need for a provider, a recipient as well as information and communication devices. They also specify that the outcome of its use "can be a benefit, service or acquisition of property" (Kvasnicova et al., 2016, p. 193). The Definition of a digital service by Pakkala and Spohrer (2019, p. 1886) shares the feature of fully relying on a technical system similar to the previous ones. But they differentiate how a specific result for a recipient is created. Namely through "digital Information, Computing, Communication and Automation Technology (ICCAT) based system[s]" (Pakkala & Spohrer, 2019, p. 1886). However, they gave no specification on how digital service are typically accessed. Online service, as a designation, has not received a definition as the previous terms have (Meszaros & Buchalcevova, 2017, p. 301). It is probably due to the reason that it is usually used synonymously or as an umbrella term for the aforementioned terms (Baida et al., 2004; Kvasnicova et al., 2016). It is mostly used to group social media, communication (messaging and video conferencing), music and video streaming, search engines, online gaming, cloud storage, etc. (Cook et al., 2017, p. 18; Kutani et al., 2021, p. 10; Obringer et al., 2021, p. 1; Schmidt, 2011, p. 30).

To boil it all down in the most basic sense: an online service is a service served over the Internet. For the purposes of this thesis, however, a final distinction must be made. Following the definition of Kvasnicova et al. (2016), online-shops and online-marketplaces, like Amazon or eBay, are exempt from the idea of an online service, since they sell physical good, still requiring help of human workers. However, services providing digital goods to consumers, like iTunes (music) or Steam (games) do fall under the definition.

Following the review of the available definitions, an online service is defined as a service offered over the internet by a provider to a user, yielding a non-physical result, generated by a system based on "digital information, computing, communication and automation technology (ICCAT)" (Pakkala & Spohrer, 2019, p. 1886).

2.4. Measuring the Impact of ICT

After defining what constitutes an online service, the next step is to look at how this affects the world in practice, in terms of environmental sustainability. Since this thesis looks at online services from a consumer perspective of daily use, the development process and acquisition of resources for an online service are not explicitly considered. As explained in the previous chapter, providers use ICT systems to make the service available to users, so there is no doubt that at least one server and one client device are needed, which of course consume electricity. Yet, this does not exactly represent the journey that a request to an online service goes through when sent by a user. In order to calculate the energy consumption and estimate the resulting carbon footprint, the majority of researchers categorize the energy consuming components into three groups: data centers, data transmission and end-user devices (Ericsson, 2020, p. 7; Hintemann and Hinterholzer, 2020, p. 3; Kamiya, 2020; The Shift Project, 2019). The most prominent example used in recent times for calculating the energy usage of an online service has been video streaming, not only because of its popularity among consumers, but also because it is said to be a considerable contributor to global energy consumption (Marks et al., 2020, p. 2; Widdicks et al., 2019, pp. 1–2). Plenty of methods exist to quantify the energy use, not only because other scopes are set, but also because factors change over time (Coroamă, 2021, pp. 14–16; Marks et al., 2020, p. 3). For this very reason, a number of sources were compared to determine a plausible range of values for the energy impact of video streaming to use as a base for the interviews (see Chapter 3.3).

The calculations for the kW/h for one hour of streaming in Table 1 are based on the methods developed in each respective paper. The energy consumption for one hour of video streaming has been set into comparison with the time an LED bulb could be powered using the same amount of energy, i.e., the same number of kW/h. To get the time in minutes, the kW/h of the bulb is divided by the kW/h of the stream and then multiplied by 60. The LED light bulb was chosen as the tool of comparison because its energy consumption is easier to grasp and no knowledge of the interpretation of the unit kW/h is required. Therefore, more minutes for the bulb mean that the online service is less resource intensive. Values range from less than a minute to about half an hour. They vary considerably, not only because different end devices were taken into account for the calculation, but also between one and the same type of device due to different assumptions about their consumption. The variance of energy consumption is easily explained as bigger screens use up more energy, Ericsson (2020, p. 11) even goes as far as to say that the end-user device is the only variable having the biggest impact, labeling data transmission and data centers as constant. Kamiya (2020) however points out that the device itself is not the only factor, as the video resolution and the type of network connection (wired, wireless, mobile) also have an effect. This results in an opposite idea of the distribution of energy consumption, Hintemann and Hinterholzer (2020, p. 4) think that the data center and network infrastructure have the bigger influence, because their energy consumption scales with the resolution of the requested video stream, as it requires more computing resources.

On a higher level, the calculation is additionally influenced by three essential principles: Moore's Law, Koomey's Law and Jevons' Paradox. Moore's Law describes the doubling of the number of transistors on a microchip in about every two years (Moore, 1965, 1975). This effective doubling of processing power per chip, or improvement in efficiency, is by some stated as a potential driver of the decrease of energy consumption by compute devices (Coroamă, 2021, p. 12; Hintemann and Hinterholzer, 2019, p. 4; Lange et al., 2020, p. 5). Others argue that this only holds true if the number of compute units stays the same and no additional servers are deployed to serve the demand (Schiettekatte, 2021, pp. 24–25). In addition, too short a replacement cycle of more efficient hardware can also have a negative impact on the environment, because the old hardware needs to be disposed of (Ericsson, 2020, p. 6). Koomey's Law builds upon the implications of Moore's Law and describes that the energy consumption of a compute unit is halved roughly every one and a half years (Koomey et al., 2011). This is brought up by several researchers as an indicator for a declining trend in energy consumption (Kamiya, 2020; Lange et al., 2020, p. 4; Maxime and Jean-Noël, 2020, p. 25).

However, some researchers point out a rebound effect that mitigates the effect of Moore's and Koomey's Law (Cook et al., 2017, p. 18; Lange et al., 2020, p. 5; Marks et al., 2020, p. 2; Santarius et al., 2020). It is also known as Jevons' Paradox, which states that technological progress that enables more efficient use of a commodity ultimately leads to an increase in the use of that commodity, rather than a decrease (Jevons, 1866). A consumer centric example of that is the market for TVs. Display technology has become more energy efficient over time, but the screen size increased as well, meaning that consumers are inclined to buy bigger TVs, making energy savings negligible (Ericsson, 2020, p. 12). Similarly, an increase in resolution, with the move from HD to 4K, can make past savings obsolete (Hintemann & Hinterholzer, 2020). Thinking further, the 'smartification' of appliances, from the already established smart TV to the smart fridge, introduces more internet connected devices to the 'traditional' roaster of personal compute devices, like computers, smartphones and tablets (Ericsson, 2020, p. 8). Those may not be influential on the global electricity use now, but might be in the future (Andrae & Edler, 2015, p. 119).

Having looked at the concerns regarding the pure energy consumption of an online service, the actual carbon footprint resulting from said consumption is up next. Ericsson (2020) developed a term for it, the *digital carbon footprint*. With

Table 1: Streaming Comparison

Author	Author background	Device and video quality	kW/h for 1h of streaming	Minutes equiv. to hour of LED Bulb
Ericsson (2020, p. 11)	Swedish ICT Provider	Smartphone HD	0,0180	18,33
		Laptop HD	0,0450	7,33
Hintemann and Hinterholzer (2020, pp. 5–6)	Borderstep Institute for Innovation and Sustainability	Smartphone HD	0,2200	1,50
		65" TV HD	0,3200	1,03
Kamiya (2020)	International Energy Agency (IEA)	50" TV 4K	0,1521	2,17
		Laptop HD	0,0462	7,14
		Smartphone Auto	0,0102	32,35
The Shift Project Materials (2019)	French Thinktank	Smartphone HD	0,6786	0,49
		Laptop HD	0,6912	0,48

Note. Minutes of video streaming equivalent to lighting a LED Bulb with 5.5W for an hour.

that said, their presentation of the term is oddly similar to how the oil company BP (British Petroleum) popularized the *individual carbon footprint* in the first place (Kaufman, 2020). This is mainly the case because mentions of the digital carbon footprint in the original white paper by Ericsson (2020) predominantly come up with 'your', as in 'the consumer', in front of it. As it appears, BP's marketing campaign was intended to deflect responsibility from themselves and their oil business and place it on the shoulders of consumers (Kaufman, 2020). The digital carbon footprint feels like it tries to hold consumers accountable for their online service consumption in the same way. This must be looked at critically, because, as stated previously, researchers are not on the same page on whether end-user devices have the biggest contribution to the whole energy consumption.

Nonetheless, discussing the concept of the digital carbon footprint has its relevance, namely in the way the actual energy consumption is converted into grams of emitted CO_2. Many researchers use the average global *energy/electricity mix* to model the carbon emissions of an online service. Sources for that can either be the International Energy Association (IEA) or national government institutions. The energy mix describes the blend of produced energy (e.g. energy from fossil fuels, nuclear, solar, wind, etc.), each having a different level of CO_2 emissions within a given electrical grid. Because of that, researchers emphasize that the actual carbon footprint is highly reliant on the source of the electrical energy and whether it is renewable or not (Cook and Jardim, 2019, p. 8; Maxime and Jean-Noël, 2020, p. 23; Obringer et al., 2021, p. 2). When an energy mix is claimed to be 'green', it also needs to be considered what the label means in that context, because some might count nuclear energy or natural gas as environmentally friendly while others might not (Cook et al., 2017, p. 38).

So for consumers to "use ICT services that help to reduce carbon emissions" (Ericsson, 2020, p. 15), they need to be aware and get reliable information about factors like screen size, resolution, energy mix, and the meaning of 'green' labels, to make an informed decision.

2.5. Bullshitting and Greenwashing

Both bullshitting and greenwashing seem to be words that are somewhat similar in their usage and meaning. One would intuitively use them to point out disbelieve in a statement or fact. With that said, 'calling out the bullshit' is a colloquial term used to criticize an opposing opinion in politics, journalism, and social media, that, through its frequent use, lost its rude and vulgar connotation in today's society (Christensen et al., 2019, p. 1588). The omnipresence of information, the constant bombardment of marketing, and the emergence of fake news could lead one to suggest that "[...] we live in an age of bullshit" (Stevenson, 2021, p. 86). By contrast, calling something greenwashing specifically refers to doubtful marketing strategies and dishonest advertisement campaigns in the context of environmental sustainability and environmental friendliness (Idowu et al., 2013, p. 1321). But greenwashing also appears in annual reports, sustainability reports, and on websites of corporations (Lyon & Maxwell, 2011, p. 9).

2.5.1. Greenwashing

The environmental activist Jay Westerveld first used the term when he tried to find a name for an issue he noticed while he stayed at a hotel (Orange & Cohen, 2010, p. 30). There were signs in every bathroom advising guests to reuse towels as a means to help save water. But when Westerveld investigated he was baffled, that the well-intentioned conservation of water, was a cost-saving measure to benefit the

business. A popular definition based on Westervelt's observation describes greenwashing as "the act of misleading consumers regarding the environmental practices of a company or the environmental benefits of a product or service" (TerraChoice, 2009, p. 1). Lyon and Maxwell (2011) put further emphasis on withholding information by defining it as "the selective disclosure of positive information about a company's environmental or social performance" (p. 9) resulting in an overly positive image of the company. Some companies are even willing to invest more money in the promotion of green practices than actual sustainability projects (Idowu et al., 2013, p. 1318).

Since a commitment to sustainability may require major changes to a company's business model, organization, or investments, greenwashing can be motivated by a company's fear of change while still meeting stakeholder expectations (Sands & Morison, 2020, p. 3). But concealing unsustainable or even dangerous practices comes with a tradeoff (Sands & Morison, 2020, p. 1). Companies can receive major backlash from the public or especially environmental activists (Jong et al., 2020, p. 68; Lyon and Maxwell, 2011, p. 4). Correspondingly, the act of accusation plays a major role in the exist-ence of greenwashing, otherwise, instances of greenwashing would be undetected (Seele & Gatti, 2017, p. 248). This introduces the risk that actions could be falsely ac-cused of being greenwashing, since full transparency cannot always be provided and unrealistic expectations can be made (Seele & Gatti, 2017, pp. 244–245). Nonetheless, information published by companies must always be critically reflected, for example, reports that are self-published are not trustworthy because they are not reviewed by independent third parties (Jones, 2019; Sands & Morison, 2020).

2.5.2. Bullshitting

When bullshit has been spotted, it is inherently seen as something bad (Luks, 2017, p. 86). Therefore, some say that uttering bullshit should always be avoided (Frankfurt, 2006; Spicer, 2018). In spite of the general 'avoiding' attitude toward bullshit, Luks (2017) points out that "without bullshit, there may be no discourse at all" (p. 90). In his view, the act of bullshitting can be utilized as a tool that contributes to discourse. Moreover, it can be used as a lubricant for stagnant discussions (Luks, 2017, p. 89).

Before discussing the concept of bullshitting in detail, it is important to know the four termini used to describe the bullshit process and its participants. First of we have the entity of the (1) *bullshitter*, they are the source for instances of bullshit (Frankfurt, 2005). Thus, (2) *bullshitting* is the name for the act of producing bullshit and (3) *bullshit* is correspondingly the product of bullshitting. During research on the reception of bullshit, the (4) *bullshitee* emerged as a later addition to the bullshit terminology (Čavojová et al., 2020; Pennycook et al., 2015). The bullshitee sits at the receiving end of the bullshitting act and is in a sense targeted by the bullshitter.

On a high level, bullshitting can be described as a communication practice with "the lack of connection to a concern with truth" (Frankfurt, 2005, pp. 33–34). However, a more in-depth look reveals, that there are two lines of research concerned with the concept of bullshitting, both having their origin in philosophy. Frankfurt (2005, p. 34) emphasized the utmost importance of the respect for truth in his definition of the concept of bullshitting. The fact that a bullshitter does not care for truth is more important than the fact that they do not speak the truth in the first place. In fact, in some situations, a bullshitter may be telling the truth because he may unwittingly report true information, not caring to verify its truthfulness (Frankfurt, 2005, p. 50). This is also what differentiates the bullshitter from a liar. In order to construct a lie, a liar needs to be well aware of the truth (Frankfurt, 2005, p. 51). They must be familiar with the true state of affairs to create an opposite set of facts, which in the best case do not contradict and compromise their position. Despite this important difference, the liar and the bullshitter have similar goals. They want to appear believable, represent their agenda credibly, or emerge as a winner from a dicey situation (Petrocelli, 2018, p. 250). Unfortunately, this makes it difficult to tell a liar and a bullshitter apart, given that one notices to be deceived at all. Frankfurt further specifies the motivation of the bullshitter: it is not about successful deception at the level of the communicated content, but about improving the bullshitter's self-presentation and the effectiveness of the manipulation (Frankfurt, 2006, p. 3).

That being said, the fact that the goal of bullshitting is to deceive others does not always have to be true (Carson, 2016). As the bullshitter is the center of Frankfurt's concept of bullshitting, he assumed that the bullshitter is always the communicator of bullshit. Cohen (2002) argues that bullshit can emerge without the intent of the bullshitter. People can unintentionally spread bullshit when they do not pay attention while sharing facts from third parties. Cohen (2002, p. 8) would describe these people as "mere victims" of bullshit, innocent and with no ill intent. In contrast to Frankfurt, Cohen (2002) tries to explain bullshit only as the product and not from the view of the bullshitters mindset. As a consequence, his definition does not refer to the motivation of the bullshitter. For Cohen (2002), bullshit in itself is characterized by vagueness and ambiguousness, its truthfulness is inherently difficult to determine. This is what he calls 'the unclarifiable unclarity', the essence of bullshit. This is different from Frankfurt's essence, which can be described as 'the indifference with truth'. Unclarity can be achieved by changing the structure of a sentence to a point where a clear meaning cannot be made out or by embedding it in a context where defining factors for its truthfulness are not present (Cohen, 2002, p. 6).

Given that there are two prevalent definitions of bullshitting, the question remains whether they are competing or even in some ways contradictory. When taking another look at the papers dealing with bullshit, it is apparent that the majority of researchers recognize both definitions and use them in combination (Christensen et al., 2019; Petrocelli, 2018; Spicer, 2018; Stevenson, 2021). To Cohen (2002, p. 1), Frankfurt's definition is seen "as bullshit of a different kind", with emphasis, that 'Frankfurt-bullshitters' can pro-

duce 'Cohen-bullshit' (Cohen, 2002, p. 8). It can be concluded that the concept of bullshitting and the concept of bullshit are complementary approaches, with Cohen's definition being the definition of bullshit and Frankfurt's definition being that of bullshitting.

2.5.3. Similar Characteristics of Bullshitting and Greenwashing

Having explained the defining characteristics of greenwashing and bullshitting in detail, the common features of both concepts will now be highlighted. Greenwashing can be categorized as a lie because it involves sharing „false[...] information" (Idowu et al., 2013, p. 1321), but lying is not seen as one of the core aspects of greenwashing (Jong et al., 2020, p. 40). Describing it as disinformation does not encompass the whole nature of greenwashing, it is more precise to say that it only takes the positive information away from the overall context to mislead the consumer, who lacks background knowledge (Lyon & Maxwell, 2011, p. 8). With that said, greenwashing as a practice is separate from lying, just as bullshitting is different from lying as stated by Frankfurt (2005, p. 46).

Sharing "deceptive information" (Idowu et al., 2013, p. 1318), "misleading consumers" (TerraChoice, 2009, p. 1), and "misleading [...] individuals who lack background information" (Lyon & Maxwell, 2011, p. 8) is regularly associated with greenwashing. Creating vague and non-transparent statements (Cook et al., 2017; Lyon & Montgomery, 2015; TerraChoice, 2009) and selectively disclosing information (Lyon & Maxwell, 2011, p. 5) are practices that coincide with Cohen's (2002, p. 6) essence of bullshit.

Lastly, greenwashing and bullshitting share a similar aspect in their motivation. Both try to hide unfavorable aspects, be it ignorance for bullshit or damaging the environment for greenwashing, to deceive their target audience into believing their manufactured self-presentation. With greenwashing, this is done by painting an "overly positive image" (Lyon & Maxwell, 2011, p. 9) about "an organization's environmental strategies, goals, motivations, and actions" (Idowu et al., 2013, p. 1318). While the bullshitter does it in any case, they want to come out on top without being held back by the lack of information. In summary, the concepts of bullshit and greenwashing can be used as a tool to critically analyze corporate claims and evaluate genuine commitment to sustainability.

3. Research Method

The following chapter explains and reasons the choice of data collection method used for this thesis. This is followed by an explanation of the selection of interview partners, the structure and composition of the interview guide and the methodology of transcription. Finally, to ensure "qualitative rigor" (p. 1) the choice of the analysis method is discussed, and it is outlined how the method is applied to the interview material (Gioia et al., 2013, p. 23).

3.1. Data Generation: Semi Structured Interviews

Given the existing research gap and the consequent small body of research, as well as the rather quantitative orientation of the thematically related papers, an explorative study is required. When problems or phenomena have not yet been examined, exploratory research is often the choice to enter the field, seeking to grasp the scope of that phenomenon or problem and to develop initial ideas and perhaps initial explanations about it (Bhattacherjee, 2012, p. 5). For qualitative research in general, interviews are a popular method to get access to "opinions, attitudes, experiences, processes, behaviors, or predictions" (Rowley, 2012, p. 261) from people in the role of a "citizen, user, consumer or employee" (Rowley, 2012, p. 260). As a specific type of interview, semi-structured interviews are often an appropriate means of data collection for the purpose of exploratory research (Alsaawi, 2014, p. 151; Gioia et al., 2013, p. 19). Being in-between a strictly predefined structured interview and an unstructured interview, sometimes only containing key topics instead of formulated questions, semi-structured interviews allow to ask follow up questions, change the order, skip parts, or ask new questions, to shape the interview around the interviewees' contributions to get the most out of them (Oates, 2006, p. 188). Interviews create a situation for people where they are able to "speak their minds" (Oates, 2006, p. 188) in a very nuanced way, which cannot be captured by a questionnaire with checkboxes or perhaps text input fields. There is no filter toward the researcher, as when a respondent cannot decide between answer choices and takes which fits best or has difficulty composing his answer in written text. An interview should be like a conversation, perhaps not as free-flowing as an informal one between friends, but it should give the interviewee room to elaborate their answers (Oates, 2006, p. 186). However, questionnaires have a significant advantage over interviews, it is much easier to gather a large number of participants, and therefore the data can be more reliably generalized (Rowley, 2012, p. 261). Nevertheless, generalization is not at the forefront of exploratory research when it first seeks to lay the groundwork. Last but not least, the opportunity to collect quantitative data should not be wasted when the decision to gather qualitative data has already been made (Bhattacherjee, 2012, p. 41). To take advantage of this opportunity, interview questions are posed in the form of *Likert* scaled questions. This allows to further contextualize the interviewees' answers with open-ended questions as well as responses by other interviewees (Frels & Onwuegbuzie, 2013, p. 188).

3.2. Selection of Interview Partners

As stated in the research question, the target group for the interviews is young adults. This group has been chosen because the proportion of internet users among young people is much higher than in all other age groups (ITU, 2021, p. 5). Thus, people in this group are more likely to engage with online services. For pragmatic reasons, as this is a bachelor's thesis, the group of possible interview partners is further

restricted to students. In order to answer the research question of whether the presence of sustainability awareness or IT knowledge changes the evaluation of sustainability of online services, the target group is divided into two sub-groups. In each case students with an IT background or students with a biology background. On one hand, this is following the assumption that students with an IT background can assess the sustainability of an online service by understanding the inner workings with the knowledge they acquire about IT infrastructure during their studies. On the other hand, it is generally assumed that biology students have a greater environmental awareness and a stronger sense of sustainability than the average student (Arshad et al., 2020; Robinson & Crowther, 2001).

The first approach to recruit participants for the interviews was to post announcements to social media and student forums, unfortunately no feedback at all came back from those attempts. In the same course the student representatives for each information systems and biology at the University Duisburg-Essen have been contacted to forward the announcement, both to no avail. Ultimately a mixture of snowball and convenience sampling was conducted by contacting fellow students that know other students that are interested in an interview (Oates, 2006, p. 98) This resulted in ten interested participants, eight of whom made an appointment.

The age of the interview partners ranges from 21 to 32 years. The IT group consists of students from computer science, information systems as well as teaching degree students in computer science and mathematics. The biology group has biology and applied life science students. All are between the 6th and 10th semesters, both in the bachelor's and master's degree programmes. Three quarters of the IT group are male (one quarter female), while three quarters of the biology group are female (one quarter male). The working title of the thesis was communicated to the interviewees during the initial contact.

3.3. Guideline and Conduct of Interviews

The interview guideline is designed in respect to answering the research question and is split up into five sections. The first section is used to gather demographic data, the second section includes questions regarding usage of online services, the third section has questions regarding sustainability. The main part includes the fourth section concerned with the perception of the energy usage of online services and lastly the fifth section presents a case of 'green' online services, with Google as an example. The original questionnaire in German can be found in appendix A.

In addition to collecting demographic data, the introductory questions also served to get the respondent in the mood for the following interview. The questions about the field of study and semester are intended to establish the academic background of the person, which sets an expectation to their knowledge. The living situation and workplace questions are supposed to give an indication of their financial dependence.

The online services section is intended to outline the interviewees' personal experience with online services and also ask about technical knowledge about them. The former is achieved by *Likert* scaled questions supported by ad hoc questions to deepen the insight, whereas the latter is an open-ended question. The *Likert Scale* is a method to rate or measure someone's attitude toward a defined topic (Likert, 1932) This is usually done by providing a scale of four or more points to let the respondent choose how strongly they agree or disagree with a particular statement (Croasmun & Ostrom, 2011, p. 19). Those kinds of rating scales are very popular among social scientists and are a regular part of most surveys (Allen and Seaman, 2007, p. 64; Croasmun and Ostrom, 2011, p. 19). The scales in the questionnaire are taken from Moosbrugger and Kelava (2020, p. 108). The decision fell on even scales to prevent the interviewees from assigning themselves to a neutral position, in order to rule out a *tendency toward the middle* if they have no strong opinions (Moosbrugger & Kelava, 2020, p. 109). The questions and scales are validated by presenting the interviewees with a version of the definition of online services from Chapter 2.3.

The aim of the third section is to identify the personal understanding of ecological sustainability and to follow up on how each interviewee acts sustainably. The connection to online services is drawn by directly asking the interviewees if they ever thought about sustainability in the context of online services. Similar to section two, the validity is given by supplementing their idea of sustainability with the definition from Chapter 2.1.

For the main part of the interview, the interviewees are introduced with an estimation question based on the calculations in Chapter 2.4. The question is intended to provide an overview of the extent to which the interviewees perceive the impact of streaming online video on the environment. However, the question about the reasons interviewees should give for their estimation is more important than the estimation question is, as it serves more as a vehicle. The question is intended to repeatedly check whether the respondents have ever thought about the sustainability of online services, but in contrast to the second section, not directly, but indirectly. It is also meant to give assess their IT knowledge.

Finalizing the main part and the overall interview is the section concerned with the interviewees' perception of bullshitting by companies using Google as an example. First off, respondents are asked about their knowledge of or experience with Google's actions toward a more environmentally sustainable business. After that, they are shown a commitment from one of Google's environmental reports. "Offer 1 billion people new ways to live more sustainably by 2022 via our core products." (Google, 2020). The commitment's focus is to enable customers to make decisions beneficial to the environment with the help of Google's services. The example given to the interviewees, to better understand the practical execution of the commitment, is listing the carbon footprint of a commercial flight next to the search results as mentioned in Chapter 2.4. The interviewees are then asked to give their opinion. Afterward, a statement from an interview from the Hessische Rundfunk with a press spokesperson of the German branch of Google is quoted to the interviewees. It states

that Google is not responsible for any carbon emissions that result from the transmission of their services over the global network to their customers (Rutsch, 2020, 16:48) (The original German quote can be found in the questionnaire in the appendix). The interviewees are then again asked to state their opinion. When conducting the interviews, an additional question was asked at the end that arose in the first interview and was not originally in the interview guide. The interviewees were asked whether they thought it would be useful for politicians to regulate sustainability claims.

The appointments for the interviews were assigned via Doodle and then conducted by video call via Zoom. The survey period extended from 03/31/2022 to 04/09/2022. The average duration of the interviews was 40 minutes. The interviewees were informed about the handling of their data in advance by mail and again before the interview. Consent for recording the conversation was given via a confirmation dialog within Zoom, otherwise, the recording would not start.

3.4. Transcription

In order to analyze the contents of the interviews, the audio files of the recordings have to be transformed into a written form by transcribing them. Transcription is a time-consuming but crucial step, as a good transcription builds the basis for conducting a thorough and transparent data analysis (Kuckartz, 2018, pp. 164–165). Transcribing one hour of recording can take up to five hours (Oates, 2006, p. 193). Luckily AI transcription tools, as provided by Word, exist, which take up the bulk of the work. Nonetheless, these tools are not 100% reliable and still require checking the transcript with the recording.

The recordings were transcribed verbatim according to the rules of the *einfaches Transkriptionssystem* [eng.: simple transcription system] developed by Dresing and Pehl (2015). Following their rules, sentence breaks, filler words, word repetitions, and stutterers are removed as far as possible. English terms are treated according to German spelling rules in upper and lower case. If necessary, statements are translated into standard German, when dialects or word slurring occur. However, grammatically incorrect sentence structures or other errors are left as such. Incomprehensible words are marked by 'unv.'. For better readability, the individual contributions are separated from each other by line breaks. Every beginning of a contribution is marked by the time stamp and the speaker. Indirect speech by the interviewee or statements in the third person are written in quotation marks or marked by a colon before the passage. Short voice dips or pitches that can be interpreted in a similar way are written as sentence endings and are therefore separated with a period instead of a comma. This serves to optimize the readability of the transcript. Statements made by the interviewee are designated by their group name (BIO or IT) and their assigned number (1 to 4). The interviewer is indicated by the initials of the author of the paper. However, some modifications have been made to the system of Dresing and Pehl (2015). Long breaks are not explicitly displayed, and short interjections are written in brackets with their respective speaker, to save space. The transcripts can be found in appendix B.

3.5. Evaluation According to Qualitative Content Analysis (Kuckartz)

Qualitative content analysis is an evaluation method that allows for systematically processing large amounts of texts (Mayring & Fenzl, 2019, p. 633). It is regularly used to analyze transcripts of interviews, where latent information is predominant. These subtle details can be found with qualitative content analysis through its primarily qualitative-interpretative approach. The procedure is strictly rule-based and thus intersubjectively verifiable, whereby the content-analytical rules are based on psychological and linguistic theory of everyday text comprehension (Mayring & Fenzl, 2019, p. 633). But as Schreier (2014, p. 2) points out, there is no 'one and only' qualitative content analysis.

After the consultation of recommended textbooks, the choice fell on the *inhaltlich strukturierende qualitative Inhaltsanalyse* [eng.: content structuring qualitative content analysis] by Kuckartz (2018). It was chosen because of its core approach of first identifying main codes deductively, and then refining them inductively into subcodes, which is ideal for guideline-oriented interviews (Kuckartz, 2018, pp. 97–98). The method has also been used in IS research, proving its usefulness in the domain (Diener & Špaček, 2021; Hacker & Bodendorf, 2017; Jost & Divitini, 2020; Wagner & Schramm-Klein, 2019) Finally, the method is supported by tools like MAXQDA, making the coding process and data management much easier.

The content structuring qualitative content analysis by Kuckartz (2018) is made up of seven phases guiding the process of the analysis. The first phase (1) is the initial reading of the text, which includes marking important passages and writing memos for salient statements. The first examination of the text prepares the basis for (2) the formation of the main codes. Usually, the main codes can be oriented by the thematic guideline of the interview. (3) Those main codes will be applied to the *unit of analysis*, which in this case is every transcript of each interview. This process of applying main codes and later subcodes is called coding (Kuckartz, 2018, pp. 35–36). When all of the material has been initially coded, (4) all text passages assigned to the same code need to be compiled together. (5) Then new subcodes are refined from those broad main codes. There can be many reasons to define a new code, but the deciding factor for refining or creating a code should ultimately always be the goal of answering the research question (Kuckartz, 2018, pp. 69–70). At this point of the process, it is also advised to start to write down the definition of each code in the *codebook* which can be found in appendix C. After the main codes have been differentiated into more specific topics, (6) a second coding of the material is carried out. It is not unusual to step back into phase five, if the is a need to define another subcode, this step of the process can be carried out many times until the material feels sufficiently coded to be analyzed (Kuckartz, 2018, pp. 110–111). It is also possible to merge codes that

were previously differentiated. The last phase (7) of the content structuring qualitative content analysis is the analysis. Kuckartz (2018) provides six different simple and complex analysis and visualization methods, but only the ones used here are discussed.

One of those would be the analysis of the correlations between codes (Kuckartz, 2018, p. 119). This can be achieved in two ways: within main codes and between main codes. Within a main code, the interrelationships that exist between subcodes are of particular interest. The main focus is on the simultaneous naming of subcodes. Here, special attention is paid to which codes are most often named together and which rarely or not at all, and, most importantly, how exactly the respondents' formulations differ within a given topic. The goal is to identify patterns or clusters of themes in the coded material. These relations can then be visualized with code relationship matrices.

Another method of analysis and visualization are crosstabs, which add a quantifying view to the qualitative data (Kuckartz, 2018, pp. 119–120). They are similar to code matrices but put more emphasis on the comparison of groups. They can be used to establish links between certain characteristics of groups, usually regarding sociodemographic data and the coded thematic statements. In the case of this thesis, the groupings are oriented by the field of study of the students. It is also possible to count and bundle the information so that a crosstab can also provide information on how frequently certain subcategories are mentioned by certain groups of respondents.

Finally, so-called summary tables can be created, in which all coded segments of a code are summarized to their core statement. All relationship matrices, crosstabs, and summary tables created in course of the analysis can be found in the results in Chapter 4.

4. Results

The evaluation of the interviews yielded three types of results: The response to items of the Likert scales, the numerical values of the estimation question, and lastly the crosstabs visualizing the coded segments according to the code system developed with the content structuring qualitative content analysis. Interesting and striking statements of interview partners are indicated by their respective ID and a timestamp referencing the transcripts in the appendix. The results will now be presented in the following chapters.

4.1. Results of Self Assessment Questions

At first glance, both groups seem quite similar in their response behavior. They are frequent users of online services and seem to be very supportive of sustainability. Many add that they use online services every day. Surprisingly, members of both groups were open to admit that they do not really pay attention to the sustainability of online services. Possible reasons for this which emerged from the interview material are presented in Chapter 4.3.

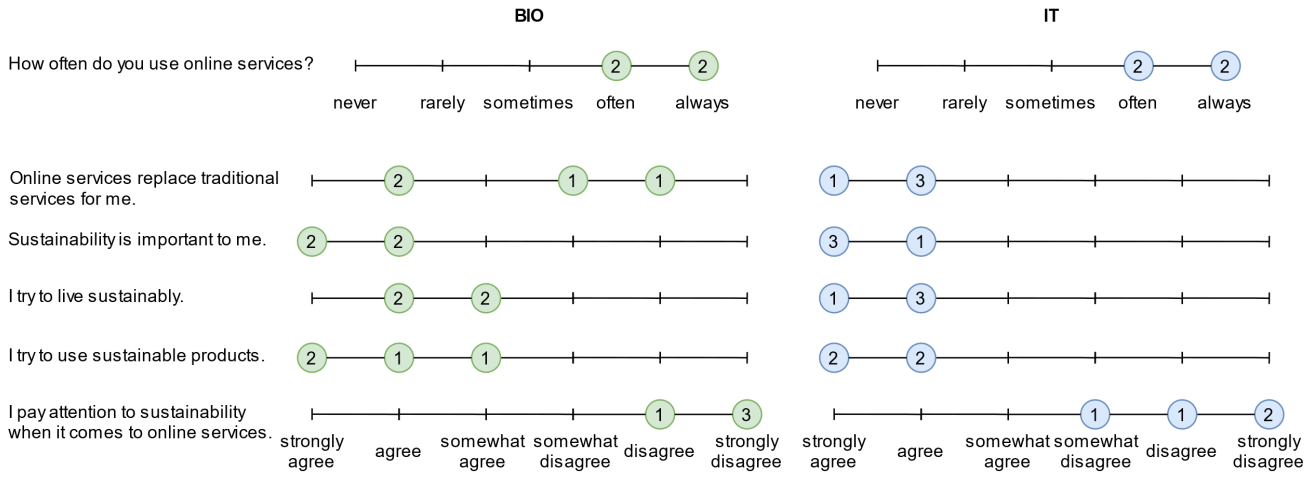
But upon taking a closer look at the results, it becomes clear that there are indeed differences in the response behavior. In the BIO group, the spread of answers to the second item is greater than in the IT group. While the members of the IT group generally agree that online services are replacing traditional services from their perspective (e.g., online banking replaces transfer slip), some BIO people disagree, although they have the same tendency to use online services regularly. It is stated that personal contact is very important and thus preferred for some types of services (BIO3, 06:59). It is also noted that some services cannot (yet) be easily implemented in a digital form (IT2, 05:36).

The members of the BIO group also seem to be more conservative in their self-assessment of being sustainable. Their item expressions do not lean as much toward 'strongly agree' as those of the IT group do. Whether the high self-evaluation of the members of the IT group compared to the rather self-critical evaluation of the BIO group is reasonable, can be judged based on their statements further looked at in Chapter 4.3.

4.2. Results of the Estimation Question

The results of the estimation question can be seen in Table 2. It should be noted that the respondents were explicitly told that they could make assumptions and express them. Most of the answers are close to the range of the realistic values previously researched (0,48 to 18,33 Minutes). BIO4, IT3, and IT4 had a very low estimate, even below the comparison range. In stark contrast, BIO3 initially estimated 80 hours but corrected its estimate to eight hours (480 minutes) because the interviewer could not suppress a surprised reaction as the estimate was far from the values researched for the interview. Nevertheless, the estimate of eight hours is still far from a reasonable estimate.

None of the BIO group members made any assumptions before giving a value. Most stated openly that they have no idea prior to guessing or when asked to elaborate (BIO1, 19:13; BIO2, 17:24; BIO3, 19:05; BIO4, 19:21). Others said that they are taking a wild guess (BIO1, 17:58) and even some are overloaded by the question (BIO3, 21:44). BIO4 guessed ten minutes at first but went down to five seconds (0,084 minutes) immediately after it was clarified that the comparison is to a LED bulb and not an incandescent light bulb. While not everyone made assumptions before giving a value, all explanations were valid in their own right. IT1 (19:48) stated that one-third (of the hour of the bulb) seems reasonable as an estimate but later added that the video resolution must influence the energy consumption (IT1, 20:26). IT3 (18:01) and IT4 (26:40) mentioned the resolution as a factor as well, while the former interviewee assumed 4K video specifically for his estimate. (IT2, 14:24) roughly outlined the IT infrastructure needed for an online service before giving his estimate. When asked to elaborate, IT3 (18:22) and IT4 (27:54) further mentioned load balancing, network bandwidth, and cooling as influencing factors for their estimation.



Note. Distribution of interviewees' responses to the question items divided by group. Numbers in dots represent frequency of the response.

Figure 1: Results of Self-Assessment Questions

Table 2: Estimated Energy Consumption of Streaming Online Videos

Interviewees	BIO1	BIO2	BIO3	BIO4	IT1	IT2	IT3	IT4
Time in min.	3 to 5	10	480	0,084	20	10	0,167	0,167 to 0,25

Comparing the two groups based on the estimated values only, they are relatively similar if the outliers are excluded. Taking a look beyond the numbers, the statements of the IT group suggest a better technological understanding. Yet, excluding the outlier, the BIO group does not guess considerably worse without the proper IT knowledge.

4.3. Results of Qualitative Content Analysis

After using the Likert scales and the assessment question to gain a rough impression of the respondents' sustainability awareness and IT knowledge, the following chapters will focus on the qualitative data that emerged from the content analysis based on the code system created. These chapters will follow the main codes of the code system, which in turn are oriented after the interview guideline.

4.3.1. IT-Knowledge

In order to assess the interviewees' IT knowledge, they were first asked to explain what an online service is to them. As shown in Table 3 on the righthand side, the majority of interviewees pointed out the most crucial aspect of online services: the need for an internet connection. Only BIO3 (01:25) opted to explain it with the help of an example instead of defining it in one sentence. Nonetheless, they named the key feature as well. In addition, all interviewees were able to give examples of online services that fall under the definition of the present work. When online shopping services such as Amazon were mentioned, it was pointed out by the interviewer that, strictly speaking, these do not fall under the understanding of online services as used in the course of the interview (BIO1, 02:57; BIO4, 02:11; IT4, 01:54). It was asked by IT1 (03:50) if Steam is also exempt from the definition but, since this service sells games as digital copies only, it also counts. All in all, however, there were no particularly unusual mentions.

Further investigation of the interviewees' IT knowledge reveals significant differences in the groups' responses. While everyone mentioned servers and end devices such as smartphones or computers in some form or another as a prerequisite for the provision and use, the members of the IT group usually went into more detail about the relationships of the components and also mentioned technical terms that are relevant in the context. Some examples can be seen on the lefthand side of the table. It has to be noted that some interviewees did not immediately start naming technical components (IT2, 08:52), as the question was not very precise in that regard, but when they were later asked to elaborate on their estimation regarding the energy consumption of video streaming, they were able to explain it precisely (IT2, 14:24).

Other prerequisites of interest were also collected as seen in the table. The interviewees with an IT background tended to mention other aspects related to the development of an online service, while the interviewees with a biology-background stated broader aspects like personnel or time, as in time available to spend using the service, as prerequisites.

Besides not naming a large number of technical components relevant to an online service, the members of the biology group tended to state openly that they are unsure (BIO4, 08:39) or simply do not have any idea (BIO2, 06:12). "Keine Ahnung" (BIO1, 09:25; BIO2, 17:24; BIO3, 10:21; BIO4, 19:21) [eng.: no idea] was a phrase used several times by interviewees with a biology background when talking about IT

Table 3: IT Knowledge Summary Table

Codes / Interviewees	IT Infrastructure	Other Prerequisites	No Knowledge	Online Service Definition
BIO1	Server, computer (as end-user device)	Time	Says to have no knowledge	Online offers, offers that have recently moved to the Internet
BIO2	Server	Personnel	Says to have no knowledge	Services that can be accessed via the Internet
BIO3	End-user device	Personnel, advertising (marketing), Time	Says to have no knowledge	Example: make an appointment online
BIO4	Server	Cooling	Says to have no knowledge	Make use of services online
IT1	Server, IT infrastructure (but does not name concrete components)	Customer service, Graphic Design	Explanation not as detailed	Service that can be accessed via the Internet
IT2	Backend server, transmission (network infrastructure), end-user device	Continuous deployment, connection to payment service (i.e., external APIs), user-friendliness		Services available via the Internet
IT3	End-user device, server capacity, load balancing, network infrastructure	User Story/user experience (requirements), platform for client application, encryption of communication, user interaction		Services provided online for different use cases
IT4	Server capacity, end-user device, internet node DE-CIX Frankfurt, private internet connection (network infrastructure)	Cooling		The main business (service) takes place via an online connection

Note. 'Other Prerequisites' refers to other prerequisites named other than those, which can be categorized as IT infrastructure.

knowledge, which in no way means that their lack of knowledge is seen negatively. They were not expected to name the same details as the IT group.

The coded segments show that the interviewees with an IT background have in fact better knowledge of IT, but as observed earlier in Chapter 4.2, the difference in knowledge among the interviewees does not indicate an obvious advantage or disadvantage in estimating the energy use on the example of video streaming.

4.3.2. Sustainability Awareness

Looking at Table 4, each interviewee was able to explain what sustainability means to them. Not all 'definitions' are exactly the same, but they all share common aspects, characterized as a practical understanding of ecological sustainability. Notably, the use of resources is often described in terms of either using them carefully or responsibly, or using resources that are described as regenerative or renewable. Reuse and recycling were also mentioned by both groups as practical expressions of sustainability. Reduction of emissions only came up within the IT group as a means to achieve sustainability, while members of the BIO group mentioned more specific aspects like animal welfare, fighting against the exploitation of people, and taking future generations into account, in terms of leaving them a healthy planet. Expressions like 'environmentally friendly' were also used to describe sustainability, which underpins the general understanding that no harm is to be done to the environment.

Table 4: Sustainability Awareness

Codes / Interviewees	Sustainability Definition	Act Sustainably
BIO1	Careful use of resources, environmental friendliness	Practices: sustainable food, sustainable clothing Perceived as difficult: has to drive a lot, for some things the time investment is too high
BIO2	Careful use of resources, mentions economic sustainability - but no further explanation, respect for future generations	Practices: sustainable food, secondhand clothing, saving electricity at home, cosmetics: shower gel and shampoo Perceived as difficult: fair trade clothing is expensive
BIO3	Responsible use of resources in manufacturing and use of products, animal welfare, no exploitation of people	Practice: buys regional products
BIO4	Recycling, environmental friendliness, conserving resources	Practices: secondhand clothing, food without plastic packaging Perceived as difficult: has guilty pleasures - not explained in detail
IT1	Use of ecologically good (renewable) resources in the manufacturing and use of products, reuse (recycling), environmental friendliness	Practices: cosmetics: packaging-free shower gel, shampoo and conditioner, toothbrush pastilles, bamboo toothbrushes, and washable cotton pads Perceived as difficult: Food with plastic packaging
IT2	Climate neutrality, environmental friendliness	
IT3	Use of regenerative (renewable) resources in the manufacturing and use of products, reuse (recycling), environmental friendliness	Practices: does not buy unsustainable food, public transport instead of car Perceived as difficult: emission compensation for parcel services, green labels
IT4	Renewable resources/energies in the manufacturing and use of products, recycling, reduce CO_2 emissions	Practices: recycled paper, wears clothes for a long time, flexitarian diet, sustainable food Perceived as difficult: buying used books

Although none of the respondents' definitions attempt to explain the concept at a high level, like scientific definitions such as the one by Ben-Eli (2018, p. 1339) in Chapter 2.1 do, everyone has a basic understanding of what is meant by sustainability and, more importantly, how to achieve it.

In order to further validate the understanding of sustainability as a concept, but more importantly, capture the awareness of sustainability, the interviewees were asked how exactly they implement sustainability in their own everyday lives. Unfortunately, the chance was missed to ask IT2 how they practice sustainability for themselves, which is why data is missing for this interviewee in the column 'Act Sustainably'. The most popular practice among the interviewees is trying to buy sustainable food whenever possible. Some are guided by labels that indicate fair trade or organic farming (BIO2, 13:24; IT1, 16:53; IT4, 23:39), others understand this to include purchasing regional products (BIO3, 16:15), and again others explicitly avoid products that have a bad reputation, for example the hazelnut spread 'Nutella', because palm oil is used in its production (IT3, 13:25). Some of the interviewees have mentioned that labels should not be blindly trusted and that it is better to pay attention to recognized labels (BIO2, 13:24; IT3, 13:25; IT4, 23:39). Others also pay attention to sustainability in cosmetics (BIO2, 13:34; IT1, 16:18). The second most popular act is buying secondhand or sustainable clothes and getting the most use out of them before buying new ones (BIO1, 15:02; BIO2, 12:49; BIO4, 10:48; IT4, 21:43). However, a hurdle was also mentioned here, relating to the fact that sustainably produced clothing is often more expensive (BIO2, 14:26). Similarly, sustainable foods were rated as more expensive, but worth spending the extra money (IT1, 39:18; IT4, 23:39). Public transport was also mentioned as a means to increase sustainability in mobility (IT3, 12:55), but some of the interviewees also admitted

that they have difficulties getting along without a car (BIO1, 12:45). In general, it was also addressed that the time required for some actions does not do justice to the preservation of sustainability (BIO1, 12:45).

All in all, respondents convey a credible commitment to the practice of sustainability in their everyday lives, however, none mentioned concrete Green IT practices, such as using secondhand smartphones or power-saving on devices at home. On top of that the lack of awareness toward online services in a direct sense, was already apparent as seen in Chapter 4.1, because everyone stated that they do not actively pay attention to sustainability in online services. A look at the sustainable activities collected indicates a lack of active efforts to achieve sustainability in the use of technology.

4.3.3. Reasons for the Lack of Sustainability Awareness of Online Services

So far, the results paint the picture that the young adults who participated in the interviews have a lack of sustainability awareness for online services. Since this is rather alarming, possible reasons and factors were collected from the interviews, which were either mentioned directly by the interviewees or emerged from the context. Table 5 groups individual reasons and factors into three groups of main codes. The first group of codes summarizes aspects that directly relate to the lack of awareness. The second group's codes collect incentives to use online services. And lastly, reasons against the use of online services were listed in the third group of codes.

With the most coded segments in both groups in total, 'Not thinking about it' is not a reason per se, but it shows that the interviewees are open to admit their shortcomings when asked. It could translate to them not having enough occasions where they are confronted with the topic. 'Powerlessness' directly refers to the belief that acting sustainably does not have an impact on the big picture, possibly lowering the intention to take action. For example, some would not feel any achievement by using an online service claiming to be green (IT2, 23:22; IT3, 25:33) because it would not make any difference (BIO1, 32:10). Segments coded with 'Unapplied Knowledge' include expressions of concern related to sustainability of online services, like cooling or energy consumption, but ultimately do not seem to have a lasting impression on the assessment, as they are followed by statements coded with 'Not thinking about it' (BIO4, 09:21; IT3, 16:56). Notably, some of this 'Unapplied knowledge' is sometimes prone to technical errors, as the assumption that the coolant used by data centers has to be 'dumped somewhere' (BIO1, 35:23), is not exactly right. In fact, most liquid cooling systems use closed-loop systems where coolant is circulated, having no need to dispose of it regularly (Ebrahimi et al., 2014, p. 627; Iyengar et al., 2012, p. 213). A very interesting reason given by the interviewees with an IT background is the fact that online services have no physical appearance and might therefore be more difficult to assess, as explained in an example by IT1 (39:18), where the quality of an organic chicken (as food) can be better judged than a service claiming to be green. Some of the members of the IT group simply stated that, as the code says, 'Privacy is preferred' to sustainability, which influences the choice of services. This will later be relevant when presenting the data for the Google example, as the same interviewees are not users of Google services, but it also indicates a clear prioritization when it comes to features of online services. While the codes 'Small User Base' and 'No Alternative' have been only coded once each, their specific problematization justify their existence. The former refers to the problem that alternative (sustainable) platforms are not attractive to new users because the initial user base is very small, which, for example, in the case of a messenger, limits the number of people with whom one can communicate on that platform (IT1, 41:55). The latter refers to the hurdle that some services might not have a sustainable alternative available. To express that there are no usable alternatives for certain online services, IT2 (26:07) uses an analogy. From the point of view of utility and energy consumption, a bicycle is a better alternative for a car than an encyclopedia for a search engine.

A big reason why the interviewees tend to prefer online services is convenience. Examples of online services creating convenience include online banking (IT1, 07:21), food delivery (IT4, 08:41), booking trips online (BIO3, 06:59), and music or video streaming (IT4, 09:14) because these offer a large selection to choose from, flexibility, and a low effort to use (BIO1, 08:46; BIO3, 02:20; IT3, 05:46). Another reason to use online services coded as 'No choice' is that sometimes one is simply forced to use them. A prominent example by the interviewees is that university documents or learning material can only be obtained via the university's online portals (BIO2, 03:44; BIO3, 07:46; IT4, 05:43). Interestingly, the interviewees do not express if they see this circumstance as a bad thing, which thematically fits in with thecode 'Ubiquity'. The interviewees say, that online services have already become part of their everyday life and that a lot can only be done via the Internet, which makes online services quietly and steadily omnipresent (BIO2, 15:24; IT2, 05:36; IT3, 04:15).

Finally, some concerns were raised about the use of online services, outside of sustainability. As already noted, online services have no physical form and, according to the definition used in this thesis, require as little human intervention as possible. BIO3 (06:59) stresses this as a downside, as they would prefer personal contact, as in face-to-face consultation, where guidance or advice is valued by the customer. They also would not like to have everything provided as an online service because to them these have the potential to replace jobs of the people, who would have performed this service previously (BIO3, 06:59). Interestingly, the convenience of an online service is also mentioned in the same paragraph, suggesting that it is seen as a trade-off.

4.3.4. Attitude Toward Google

During the interview, it has already become clear that it cannot be assumed that every participant uses Google as their preferred search engine. Therefore, it was also surveyed

Table 5: Reasons for Poor Awareness

Group	**BIO**		**IT**	
Mentions	Once per int.	Total	Once per int.	Total
Reasons to Disregard Sustainability of Online Services				
Not thinking about it	4	10	4	8
Powerlessness	3	4	3	3
Unapplied Knowledge	2	4	2	2
Immateriality	0	0	2	3
Privacy is preferred	0	0	2	2
Small User Base	0	0	1	1
No Alternative	0	0	1	1
Reasons to Use Online Services				
Convenience	3	5	3	8
No Choice	2	3	1	1
Ubiquity	1	1	2	2
Reasons to Not Use Online Services				
No Personal Contact	1	1	0	0
Replacing Jobs	1	1	0	0

Note. 'Once per Int.' shows the appearance of the code counted once per interview. 'Total' counts every instance of a coded segment in all interviews.

which search engine was used and why. This resulted in the following data, which can be seen in Table 6.

Among all interview participants, the majority of interviewees use Google as the preffered search engine. It has to be pointed out, that two of these people were not exactly sure at first if they use Google as a search provider. BIO3 (26:01) said that they are using Safari on a Mac and therefore not using Google Search, it was made clear to them that they are probably using Google because it is the default option in Safari. Similarly, IT1 (41:30) stated that they directly type the search query into the address bar. They realized soon after that, that Google would probably be the default search provider in that case too. BIO2 (19:58) and IT3 (23:09) did not explicitly state that they do not use Google by the time the interview section has been introduced. (BIO4, 15:23) stated they use Google, but should use Ecosia, as the 'sustainable alternative' to Google. Nevertheless, they stay with Google out of convenience and shy away from switching.

The two that have specified to use DuckDuckGo, said that they want to protect their privacy which in this case is more important to them than sustainability (IT2, 17:56; IT4, 25:11). However, IT4 (25:55) added that they occasionally use the Google Drive cloud storage service with a throwaway account.

Out of the four people who know Ecosia as a sustainable search engine, there is only one person who actually uses it. Those who know about it did not seem particularly convinced of it either. Ecosia was described as a search engine that 'supposedly' plants trees for each search query (IT3, 25:33). IT1 (39:18) said they know Ecosia but are not familiar enough with it to declare if the company behind the search engine really plants the trees as they claim. Even the one that uses Ecosia expressed, that Ecosia at least advertises that they act sustainably (BIO1, 21:28). They knew about the rough number of trees that have been planted so far but disclosed that they did not research it themselves, but got the information from their professor (BIO1, 27:40).

After clarifying which search engine they prefer, the interviewees were asked, what they knew about Google's sustainability efforts. Most of those interviewed said they did not know anything about it (BIO2, 19:52; BIO3, 25:47; BIO4, 26:45; IT1, 25:29; IT4, 30:40). Hardly anyone was aware of the slogan "CO_2-neutral seit 2007" [eng.: carbon neutral since 2007] at the bottom of the German desktop version of Google Search. Some were even surprised that Google is committed to sustainability (BIO4, 27:10). Others said that they only know negative headlines about Amazon and their data centers and therefore assume that Google, as a big tech company, cannot be any better (IT1, 26:06).

The opposing code to 'No Knowledge [...]' is called 'Some Knowledge about Google's Efforts' with emphasis on 'some', because the statements were very broad or not exactly correct. Of course, comprehensive knowledge on the subject cannot be demanded, but only one example would

Table 6: Preferred Search Engine

Interviewee / Preferred Search Engine	BIO1	BIO2	BIO3	BIO4	IT1	IT2	IT3	IT4
Google		X		X			X	
DuckDuckGo						X		X
Ecosia	X							
Aware of Ecosia				X	X	X	X	
Unsure			X		X			

Table 7: Knowledge about Google's Sustainability Efforts

Interviewee / Knowledge	BIO1	BIO2	BIO3	BIO4	IT1	IT2	IT3	IT4
No Knowledge	X	X	X	X	X			X
Some Knowledge						X	X	

not suffice for basic knowledge if one presents oneself as particularly committed to sustainability. As an example, IT2 (17:07) heard that Google has started to build data centers in the sea to reduce electricity costs for cooling. After asking a follow-up question, it was specified that Google submerged tanks with servers inside, which were presumably deposited off the coast of Scotland. But it turns out this was not Google after all. The venture was a pilot project by Microsoft (Cellan-Jones, 2020) therefore it must be a mix-up.

Now that the interviewees' 'basic attitude' towards the company has been presented, their statements on Google's commitment and the spokesperson's statement can be addressed. This was achieved by looking at the relationships between the codes describing the attitude toward Google and the 'basic attitude' from before the commitment and the statement were presented to the interviewees. A relationship exists if a code from the y-axis and the x-axis of the table occurs in the same transcript, i.e., they were mentioned by the same interviewee. The four tables 8 to 11 have the codes for the 'Attitude toward Google' on the left and the codes indicating their preferred search engine and knowledge about Google's sustainability effort on the top. It is important to add that the codes on the left differ between the context of the commitment and the statement, as new aspects were added by the interviewees when the statement was disclosed. The results for the BIO group and the IT group are displayed in separate tables because of size constraints. The maximum value in the top line indicates the maximum number of relations a code can have, since there can only be as many relations as there is people coded with the code in the top row (see Table 6 and Table 7).

Looking at the tables 8 and 9, it is apparent that the most common code is 'Better than nothing' with no definite affiliation to a group or a 'basic attitude'. For example, the participants say that the commitment is well-intentioned, but the effect of it needs to be proven (IT2, 21:53; IT3, 24:42). The option for choice is welcomed, but it is speculated whether the implementation of the commitment has only a small effect or any effect at all (BIO2, 22:53; BIO4, 28:23). All in all, none seem particularly enthusiastic about the commitment and see the exemplary implementation of it (see Chapter 2.1) as a 'nice to have' rather than a game-changer. The overall skepticism against Google and its actions toward sustainability seems to be more apparent within the IT group. As in 'we need to question this' (IT3, 13:58), 'is it really like that' (IT1, 32:12), or 'what do the numbers actually mean' (IT4, 31:59). Concrete criticism however is comparable in number between both groups.

'Carbon Offsets' were only mentioned by one interviewee in the IT group out of all participants. They assumed that Google achieves its carbon neutral status by using those instead of actively reducing its emissions (IT1, 35:10). 'Conflicting Interests' were also mentioned by only one respondent from the BIO group. What they saw as a conflict of interest in Google's implementation of the commitment is, that the carbon rating and advertisements by third parties on their platform could compete, potentially impacting the revenue stream for Google (BIO1, 24:05). To them, this would compromise the implementation of the commitment. Both IT1 (39:18) and BIO4 (10:48) call Google's slogan and their implementation of the commitment greenwashing, and even bullshit. They describe it as misleading advertisement and as a scam [ger.: "Masche"] alike (BIO4, 28:23). Although not

Table 8: Attitude toward Google (Commitment/BIO group)

Attitude toward Google	Aware of Ecosia (max. 1)	Google (max. 2)	Unsure (max. 1)	Ecosia (max. 1)	No Knowledge (max. 4)	Total mentions
Better than nothing	1	2	1	1	4	7
Carbon Offsets	-	-	-	-	-	0
Conflicting Interests	-	-	-	1	1	1
Greenwashing	1	1	-	-	1	3
Lack of Transparency	-	-	-	1	1	1
Offloading to Consumers	1	2	-	-	2	3
Skepticism	-	-	-	1	1	1

Table 9: Attitude toward Google (Commitment/IT group)

Attitude toward Google	Aware of Ecosia (max. 3)	Google (max. 1)	Unsure (max. 1)	DuckDuckGo (max. 2)	No Knowledge (max. 2)	Some Knowledge (max. 2)	Total mentions
Better than nothing	3	1	1	2	2	2	5
Carbon Offsets	1	-	1	-	1	-	1
Conflicting Interests	-	-	-	-	-	-	0
Greenwashing	1	-	1	-	1	-	2
Lack of Transparency	2	-	1	1	1	1	3
Offloading to Consumers	2	-	1	2	2	1	3
Skepticism	2	1	1	1	2	1	5

directly related to Google, it was noted under 'Lack of Transparency' that BIO1 (26:48) does not think that the information from Ecosia, on how many trees they planted, is realistic. They said that they could 'theoretically' get information on that, but the wording makes it sound more like they have not done so yet. The information released by companies is generally seen as ambiguous (IT1, 39:18; IT2, 20:14). The code 'Offloading to Consumers' is present in both groups again. Some interviewees expressed outrage, because the commitment seemed like Google was trying to transfer its responsibility to the customer while taking credit for the sustainability efforts (IT1, 30:06; IT4, 33:26). IT2 (21:53) even links it directly to the 'scandal' about BP's individual carbon footprint discussed in Chapter 2.2. Members of the BIO group call the commitment an excuse for Google to call itself sustainable or wishful thinking that people do make the more sustainable decision thanks to a Google service (BIO2, 24:35; BIO4, 31:50).

As already mentioned, the codes in the following tables (Table 10 and Table 11) are slightly different from the ones before. 'Better than Nothing' and 'Carbon Offsets' did not come up again in the context of the statement by the Google spokesperson. However, the codes 'Intolerable Behavior', 'Statement is Understandable', and 'Company has the power to change' needed to be defined to capture statements concerned with these aspects. Since 'Better than Nothing' is one of the few codes that can be associated positively, the overall mood shifts to the negative after the statement is presented. From all of the newly emerged codes, 'Intolerable Behavior' has the most negative connotation. The initial reaction to the statement of the spokesperson is that the image of Google is 'weak' and 'wrong', as they are not carbon neutral on all accounts (IT1, 44:09). Most interviewees agree that the route traveled via the Internet should also be factored in. Others say that since this statement is true, carbon neutrality should not be advertised in this way because it is dishonest (BIO3, 33:28, 38:14). Some even go so far as to say that this is consumer fraud, as the statement is disguised by the general public portrayal of Google (BIO2, 33:46). The code 'Company has the power to change' has two slightly different variants: on one hand it is assumed that Google, as a major player in the market, could use its power to enforce regulation in its supply chain (IT3, 30:32; IT4, 39:14). On the other hand, it is demanded that they just do so (BIO4, 37:27). While the first assertion, appeals to Google's drive, the second skips this and demands direct action. Going fur-

Table 10: Attitude toward Google (Statement/BIO group)

Attitude toward Google	Aware of Ecosia (max. 1)	Google (max. 2)	Unsure (max. 1)	Ecosia (max. 1)	No Knowledge (max. 4)	Total mentions
Company has the power to change	1	1	-	-	1	1
Conflicting Interests	-	-	-	1	1	1
Greenwashing	1	2	1	-	3	5
Intolerable Behavior	-	1	1	-	2	3
Lack of Transparency	-	1	1	1	3	5
Offloading to Consumers	-	-	-	-	-	0
Skepticism	-	-	-	1	1	1
Statement is Understandable	1	1	1	-	2	2

Table 11: Attitude toward Google (Statement/IT group)

Attitude toward Google	Aware of Ecosia (max. 3)	Google (max. 1)	Unsure (max. 1)	DuckDuckGo (max. 2)	No Knowledge (max. 2)	Some Knowledge (max. 2)	Total mentions*
Company has the power to change	1	1	-	1	1	1	2
Conflicting Interests	-	-	-	-	-	-	0
Greenwashing	1	-	1	-	1	-	1
Intolerable Behavior	1	-	1	-	1	-	1
Lack of Transparency	-	-	-	-	-	-	0
Offloading to Consumers	-	-	-	1	1	-	1
Skepticism	-	-	-	-	-	-	0
Statement is Understandable	1	1	-	2	1	2	3

ther, 'Statement is Understandable', while not overly positive, shows some interviewees to be in part considerate of Google. It is accepted that Google does not have control over the entire route that their online services take over the network (IT3, 31:58). For most, however, this consideration does not last long.

This 'truth' is then contrasted with the fact that one then cannot advertise with complete carbon neutrality (BIO3, 38:14). Wich in part leads back to the idea of 'Offloading to Consumers' (IT4, 39:14). To preserve its legitimacy it should then be Google's obligation to stop such misleading practices (BIO4, 37:27). Some said they had already guessed that there would be a catch, and they did not expect to be surprised (BIO1, 34:32; IT2, 28:03).

The tables show that the code 'Greenwashing' occurs more frequently in the BIO group after the statement has been presented, while it remains unchanged in the IT group. IT1 (45:35) consolidated their point of view by reiterating their greenwashing accusation. BIO4 (36:02) also reinstated the point, that the slogan is 'bullshit' and felt vindicated. Google's image is now described as 'contradictory', 'fake', and 'different behind the scenes' (BIO2, 30:07, 31:38; BIO3, 33:28). 'Lack of Transparency' came up more frequently in the BIO group as well, by contrast, the aspect was not repeated in the IT group. BIO1 (29:57) used the example of Amazon to explain that it is tricky to understand what actually happens when an order is placed. They criticized that those orders show up in multiple packages even though it was chosen to deliver them all at once. BIO2 (33:46) followed suit to admit that it is difficult to assess such claims, they also added that the statement allows Google to use it as an excuse if they are accused of lying. More transparent handling of the matter was demanded by BIO3 (34:42). Lastly, the codes 'Conflicting Interests' and 'Offloading to Consumers' appeared only once again each. The former was repeated in a slight variation (BIO1, 35:23) and the latter was mentioned again to underpin the point that the commitment is dishonest (IT4, 39:14).

4.3.5. Opinions on the Responsibility for Sustainability

Between the commitment and the statement part of the interview, the interviewees were asked which side, i.e., consumer or company, should take more responsibility for sustainability. This and the emergent question about political intervention provide the data for Table 12. The table clearly shows that there is no permanently assignable opinion, neither to a group nor to specific interviewees, as they sometimes express several viewpoints.

From left to right, BIO1 has the most varied opinion. To them, the main responsibility was first with the companies (28:12), then it was added that the customer is also in part responsible (28:12). After that, it was noted that companies cannot be responsible for everything (29:57) and at last it was stated that no one wants to take full responsibility (32:45). BIO2 (25:35) took a clearer stand, to them the responsibility is of equal parts. BIO3 (31:27) shared the same opinion and emphasized that everyone must do their part. BIO4 (34:01) takes the stance that companies should be made responsible, as they have not cared in the past. To them, relying on the consumers" initiative yields little to nothing. IT1 (36:27) would see the responsibility on the company's side because they are in the position of power. They added that the purchasing power of the consumer could influence a company's actions if an alternative product is available, but they still think that the contribution of the consumer is important (45:35). IT2 (24:25) argued that it is not easy to decide who should bear the responsibility, but they stressed that companies who call themselves sustainable should be held liable for it (30:52). IT3 (27:41) is the only one among the interviewees who thinks that the responsibility is with the consumer, they justify it by market mechanisms. To them, companies would move toward sustainability through supply and demand (27:58). Lastly, IT4 (35:44) would not distinguish between 'either-or' but would attribute significantly more to companies because they are the main polluters.

5. Discussion

The research question aims to address two sub-aspects, firstly the effect of sustainability awareness and IT knowledge in assessing the sustainability of online services, and secondly whether bullshit claims about the sustainability of these online services can be detected. To evaluate if the collected qualitative data contribute to answering the research question, the results are discussed in the respective chapters with regard to the two sub-aspects. Subsequently, the theoretical contribution and practical implications, as well as the limitations of the work, are addressed.

5.1. Influence of Sustainability Awareness and IT Knowledge

The results strongly suggest that it does not matter whether a higher level of sustainability awareness or IT knowledge influence the awareness of the sustainability of online services, as all participants effectively say that they do not consciously pay attention to whether an online service is sustainable or not. With that said, the sustainability awareness is similar between groups, so no direct link can be made to IT knowledge. A more detailed survey of personal sustainable practices based on the *Pro-Environmental Behavior Scale* (PEBS) by Markle (2013) might have allowed for a finer distinction, but would also have taken time away from other important parts of the interview. The PEBS would enable to measure the 'performance' of the environmental behavior rather than describing it as basic, as in the case of this thesis (Markle, 2013, p. 912). The assessment of IT knowledge was clearer, as the IT students were able to explain the background of online services more convincingly and in much greater detail. Accordingly, the statements were also technically correct when compared to the conceptual foundation of this thesis.

As previous research projects have shown, it cannot generally be concluded that Green IT will be practiced more just because sustainability awareness is increasing among the population as a whole (Koo et al., 2015, p. 75; Kurkoon et al., 2018, p. 9; Leung et al., 2018, pp. 2–3). This is also confirmed by this thesis in the context of online services, which have not yet been studied in such detail. Likewise, the interviews uncovered reasons responsible for the low awareness of sustainability concerning online services. These will now be discussed and compared to existing literature to see if the same or similar findings have already been made or if other barriers exist exclusively for online services.

The most prominent issue is the lack of knowledge that came up with the code 'Not thinking about it'. It is not to be understood as a lack of interest, but as a lack of education, as the interviewees made the impression that they were eager to know how online services can be sustainable (BIO3, 17:37). The fact that the interviewees do not practice sustainability in terms of online services, because they do not know where to start, coincides with the findings of Leung et al. (2018, p. 9), who conclude that a good understanding of Green IT increases the confidence in practicing Green IT. Kurkoon et al. (2018, p. 4) point out that 'noticeability' is also an important factor influencing the intention to use Green IT products. Noticeability is understood as the ability to recognize green products as such, based on the observation that consumers would like to use sustainable products, but most of the time do not know what qualifies as a sustainable product. This is reminiscent of how the interviewees initially did not know what a sustainable online service could be, with the exception that only a few named Ecosia as an example.

The idea of the 'Perceived Green Benefit' by Kurkoon et al. (2018, p. 4), goes a step further, as it partially requires knowledge about Green IT. The perceived green benefit describes the added value of a sustainable product that a consumer expects individually. The added value, i.e., the advantages of a Green IT product, are also crucial for the intention to use them. Besides partly relating to the code 'Not thinking about it', it is comparable to the statements coded as 'Powerlessness', as it summarizes the feeling that acting sustainably does not have an impact, thus not creating any value in

Table 12: Opinion on who has Responsibility for Sustainability

	BIO1	BIO2	BIO3	BIO4	IT1	IT2	IT3	IT4
Equal Responsibility	X	X	X		X	X		
Responsibility of Consumers/Users	X						X	
Responsibility of Companies	X			X	X	X		X

choosing a particular green service. An extreme example of that would be the statement of IT3 (25:20), who does not think that Google's services provide any sustainable benefit to them. Despite that, they still use Google for searching the web, which from the sustainability perspective goes against the point of Koo et al. (2015, p. 71), who argue that the perceived usefulness prolongs the use of green products. However, it is apparent that the reason for using Google is not primarily sustainability, but convenience.

To be swayed by convenience is another important aspect that came up in the interviews and has been discussed in previous research. Kurkoon et al. (2018, p. 4) address this with the idea of 'Resource Sacrifice', meaning that green products or services tend to be less feature-rich and lack performance in comparison to non-sustainable ones. While the interviewees did not draw a direct comparison between a sustainable and a non-sustainable service, it was very evident that they value the convenience of online services over regular services, as listed in Chapter 4.3.3, and conversely would rather not return to the 'old' way. The issues labeled as 'Small User Base' and 'No alternative' also share the sentiment of not wanting to give up something in turn for sustainability.

For the example of video streaming, Leung et al. (2018, p. 8) found that people would choose streaming providers based on cost and quality rather than sustainability when they indicated that they found it difficult to practice Green IT. Although it is not about the same properties, this finding is similar to the preference for privacy of some interviewees.

One aspect that is only briefly touched upon by previous research is the immateriality of online services. It is mentioned that "green products and services are likely to be abstract objects" (Kurkoon et al., 2016, p. 599) from the perspective of consumers, making it difficult to assess them. But given the context, which further explains the idea of noticeability mentioned earlier, the word abstract is referring to green products and services being difficult to grasp and not the abstraction of a service from the material to the digital world. Koo et al. (2015, p. 68) argue that the effects of Green IT and Green IS can actually be seen by consumers across products and services. However, the results of this thesis contradict this assertion, because the interviewees state that sustainability is not immediately apparent to them in online services. Koo et al. (2015, p. 68) further note that the usefulness perceived by the consumer depends on whether the effects of sustainability are obvious to the consumer. This might be true for the power monitoring devices looked at in their study, but online services seem to act differently in that regard. Since the interviewees struggled to make sense of the sustainability of online services, the effects are not obvious or perhaps not made obvious enough.

One last observation worth mentioning is that the interview partner who uses Ecosia does it based on a recommendation and does not really seem convinced either (BIO1, 22:03). However, Kurkoon et al. (2018, p. 9) described the influence of the social environment as insignificant, which contrasts with previous studies that assumed the opposite, as they themselves affirm. Even if this is only a single case, it is still interesting to note that this usage decision is motivated more by extrinsic than intrinsic motivation. Since all other sustainability practices seemed to be done out of conviction. Thus, social influence is credited with having some degree of impact.

Looking at the discussed aspects, it can be concluded that some influencing factors of Green IT products can be transferred to online services, but hurdles have also been identified that have not been considered so far in research. In the end, it does not matter whether sustainability awareness or IT knowledge is more strongly developed, in the end, Green IT knowledge must be promoted directly to the consumer as already concluded by Leung et al. (2018, p. 9).

5.2. Ability to Spot Bullshit Claims

The question of whether the interviewees were able to debunk Google's bullshit claims is difficult to answer. Overall, it is clear that the interviewees generally had doubts about the feasibility or usefulness of the commitment that Google has made, even though the commitment is considered good in principle. Moreover, this critical attitude develops and is confirmed or even strengthened with the unveiling of the statement of the press spokesperson. Now, this stands however in strong contrast to the conclusion that as good as nobody pays attention to the sustainability of online services and beyond that nobody really knew what Google does for sustainability at all. There is no discernible connection between the coded critical statements, and the preferred search engine and knowledge about Google's activities to promote sustainability.

This raises the question of whether this criticism and skepticism are justified at all when no one really knows what Google contributes to sustainability in practice. But it also seems obvious that this is precisely the problem. Many interview partners have mentioned the lack of transparency and

difficult comprehensibility as a point of criticism. This could explain the aforementioned lack of knowledge. However, it is unclear whether it is the lack of transparency on the part of the company or the ignorance of consumers who regularly use Google. There is a hint of this kind of ignorance because at one point it was subliminally admitted that someone could look it up if they wanted to (BIO1, 26:48). The others have not expressed themselves in this way, but it cannot be ruled out that they too act this way. This is, again, difficult to determine from the interview material ex-post. It would have required asking questions regarding the gathering of information, which was not considered beforehand.

However, it would generally be wrong to completely doubt the basic skepticism of the interviewees toward Google because critical questioning is right and important. But one can ask whether the interviewees might not be bullshitting themselves. It does not fit: they do not know what Google is doing at all, criticize them, and partly use their services anyway. It almost seems like a form of 'meta-bullshit'. Accusing Google of bullshitting for downplaying important details in marketing is one thing, but doing so on a basis of unawareness, which fits the practice of bullshitting, makes the accusation seem less genuine. Still, the bullshit was recognized, so the assessment of the interviewees is correct. Yet, as stated in the conceptual foundation, the produced bullshit can still be 'true' despite the lack of knowledge about the 'truth'.

However, it can be ruled out that the kind of bullshit in which the interviewees are interwoven is Frankfurt bullshit. This is the case because the interviewees are presumably not pursuing the motive of appearing particularly critical and environmentally aware in the interview. Otherwise, they would have claimed from the start to pay attention to sustainability in online services. The interviewees rather fell for Cohen bullshit. That is justified in the fact that many of the interview partners came back to the assumption that large companies, including Google, can only be up to no good (BIO1, 32:45; BIO4, 34:01; IT1, 26:06). This is an assertion that is often said about large companies in the vernacular, which is also confirmed by scandals from time to time, see the example of BP in Chapter 2.4. The skepticism is therefore based mostly on the assumption that large companies are bad, instead of having concrete information. Since the interviewees have no concrete knowledge about Google's sustainability efforts, it has to be said that Google does not do a good job of being convincing and trustworthy when it comes to sustainability, provided that they really do what they say. It remains surprising that the direct accusations of bullshit and greenwashing also tended to come from those who even use Google, although with such a small number of respondents this could also be a coincidence. If the previous interpretations are summarized, it can be stated that there is a discrepancy between what is done and what is said. Without the interviews, this discrepancy might not have become visible at all, because the interviewees would not have been asked about it.

It is surprising that among all the critical comments, there were also concessions to Google, some of which expressed agreement that Google has no influence. This is in contrast to the fact that Google is credited with having the power to change something. What is particularly striking here is that in some cases one and the same person expressed these points, which in itself sounds contradictory (BIO4, IT3, IT4). Yet whether this is an inconsistency in the interviewees' opinions or a misfortune in formulation lies within the scope of interpretation of the material and cannot be answered unequivocally.

Lastly, the question about responsibility for sustainability did reveal the positions of individual interviewees. However, it did not help to identify patterns in the interview material. The opinions of the interviewees on this topic are varied, as are their reasons. What can be inferred, however, is that most of them do not see the responsibility entirely in the hands of one party. Which partly reflects the obligation of the companies to publish information [ger.: 'Bringschuld'] and the obligation to collect information of the consumers [ger.: 'Holschuld'], as a result of the transparency dilemma mentioned earlier.

6. Conclusion

In conclusion, neither the level of sustainability awareness nor the IT knowledge of young adults contribute to the assessment of the sustainability of online services , as the results of the of the interview show. It has been found that interviewees have few clues about the issue of sustainability in online services. Reasons contributing to this lack of knowledge were identified and discussed. The latter part of the research question, regarding the ability to spot bullshit claims, can be partially answered but not fully explained. While the interviewees did point out issues with Google's portrayal of being sustainable, those callouts of 'bullshit' were rather based on general assumptions than on concrete knowledge. It could not be finally answered how there can be a connection that the interviewees know almost nothing about Google in the context of sustainability but are nevertheless aware of the possible bullshit. In addition, it remains unclear why some still use Google altogether.

6.1. Theoretical Contribution

On the one hand, this thesis contributes through theory building as new observations were made. Most notably, the immateriality of online services as a reason for the difficulty in evaluating their sustainability or the discrepancy between the attitude toward Google and the use of its services. From these phenomena, new hypotheses can be derived that can in turn become the basis for a theoretical explanation.

On the other hand, this thesis also contributes by theory testing to some extent. Previous hypotheses and theories from the literature that are originally applied to the quantitative approach of questionnaires, were field-tested with the qualitative approach of interviews. Both theory testing and theory building are of utmost importance to create good theories and enable "advancements in science" (Bhattacherjee,

2012, p. 4). Since the approach to answering the research question in two sub-aspects provided both expected and unexpected results, it allows for progress in both areas. Furthermore, this study adds to the existing literature on the reception of green IT products and practices by more specifically looking at the research subject of online services.

6.2. Practical Implication

The main implication of this thesis is that users of online services should be made more aware of the factors that influence the energy consumption of such a service or ICT in general. Consumers should be empowered, to look beyond the general criticism of ICT-providing companies and engage with what they claim to be doing for sustainability. In addition, they should also obtain concrete third-party information on that if possible, to separate the bullshit from the facts.

Companies can also learn from the results of the interviews. They should take care to provide information in a more accessible way to have a more convincing effect on consumers. Efforts to promote sustainability should not be misunderstood as mere marketing when it is in fact genuine.

Nevertheless, the information problem, as discussed, comes from both sides, so the information must be accessible enough to be accepted in order to reach the masses. Only in this way can sustainable development, driven and supported by consumers, achieve its full potential.

6.3. Limitations

As with any interview with pre-structured questions, it cannot be ruled out that other factors or topics not asked in the interviews may influence the results of the research. For example, a missed question remains: Would those who use Google's services continue to use them after critically discussing them in the interview, or would they perhaps refrain from doing so? It could be hypothesized that they ignore the criticism they have voiced at the next opportunity, a la 'Ignorance is Bliss'. What speaks for this is the 'convenience', as described in Chapter 5.1 but since the interviews did not cover this point in-depth, this cannot be answered with what was learned from the results. In the same vein, the phrase "guilty pleasure" (BIO4, 16:38) is also interesting but was unfortunately not discussed further in the interview. Furthermore, it was wrong to assume that everyone who could participate in the interviews is familiar with Google. Although the opportunity to gain insights from other search engine users was embraced, it could have been better integrated into the interviews if this had been considered from the beginning.

The generalizability of the results from young adults to consumers is limited due to the sample size. Simply said every student is a consumer, but not every consumer is a student. The target group only represents a section of all possible types of consumers. In addition, the interviews could very well lead to different results, if other interview partners were acquired, let alone if the age group of the field of studies were defined differently. In turn for gaining deeper insight, the individual nature of the interviewees' responses further limits the generalizability of the results.

Furthermore, the study is limited by the objectivity of the researcher, as they might have an influence on the interview (Alsaawi, 2014, p. 155). Because of that, the results could be unconsciously biased in a favorable direction. This bias was mitigated to the best of the researcher's knowledge by using the method of semi-structured interviews in which the interview guideline was reviewed by third parties to avoid leading questions. Nevertheless, this study advances the research on the topic and contributes to closing the research gap.

6.4. Future Research

Since this work has encountered some limitations in its execution, these should be taken into account in further research. Among those limitations were, as explained, missed aspects, questions, and the general issue of generalizability. To address these, researchers could design a survey based on the interview guideline and the code system to see if the same results occur in a large group of people. Especially the codes for the reasons for the lack of sustainability should be verified with a more representative group of respondents.

In addition, two hypotheses can be derived from the discussion that should be considered in particular in the future. First, immateriality affects the assessment of online services in terms of sustainability. And second, convenience influences the willingness to change the current online service usage behavior. Furthermore, the discrepancy between factual knowledge and broad criticism toward Google should be further investigated, as it is unclear whether this behavior is limited to the interview participants and how it can be explained.

The research could also be conducted from a different perspective. Instead of looking at the reception of consumers, exclusively, it could also be investigated what companies can do to be more genuine and trustworthy when actively contributing to sustainability.

Finally, it should be considered which measures can be developed to help consumers become more aware of sustainability when choosing their online services. As the present thesis exposes the shortcomings in the handling of the topic by companies and consumers alike.

References

Allen, I. E., & Seaman, C. A. (2007). Likert scales and data analyses. *Quality Progress*, *40*, "Article 7", 64–65.

Alsaawi, A. (2014). A Critical Review of Qualitative Interviews. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2819536

Andrae, A., & Edler, T. (2015). On Global Electricity Usage of Communication Technology: Trends to 2030. *Challenges*, *6*(1), 117–157. https://doi.org/10.3390/challe6010117

Arshad, H., Saleem, K., Shafi, S., Ahmad, T., & Kanwal, S. (2020). Environmental Awareness, Concern, Attitude and Behavior of University Students: A Comparison Across Academic Disciplines. *Polish Journal of Environmental Studies*, *30*(1), 561–570. https://doi.org/10.15244/pjoes/122617

Baida, Z., Gordijn, J., & Omelayenko, B. (2004). A shared service terminology for online service provisioning. In M. Janssen, H. G. Sol, & R. W. Wagenaar (Eds.), *Proceedings of the 6th International Conference on Electronic Commerce - ICEC '04* (p. 1). https://doi.org/10.1145/1052220.1052222

Ben-Eli, M. U. (2018). Sustainability: definition and five core principles, a systems perspective. *Sustainability Science*, *13*(5), 1337–1343. https://doi.org/10.1007/s11625-018-0564-3

Bhattacherjee, A. (2012). *Social Science Research: Principles, Methods, and Practices* (Second edition). Anol Bhattacherjee.

Büttner, L., & Breitkreuz, A. (2020). Arbeiten Nach Corona: warum Homeoffice gut fürs Klima ist. https://nbn-resolving.de/urn:nbn:de:kobv:109-1-15435504

Capgemini (Ed.) (2021). Sustainable IT Report. https://www.capgemini.com/de-de/news/sustainable-it-nachhaltigkeit-studie-2021/

Carson, T. L. (2016). Frankfurt and Cohen on bullshit, bullshiting, deception, lying, and concern with the truth of what one says. *Pragmatics & Cognition*, *23*(1), 53–67. https://doi.org/10.1075/pc.23.1.03car

Čavojová, V., Brezina, I., & Jurkovič, M. (2020). Expanding the bullshit research out of pseudo-transcendental domain. *Current Psychology*. https://doi.org/10.1007/s12144-020-00617-3

Cellan-Jones, R. (2020). Microsoft's underwater data centre resurfaces after two years. *BBC News*. https://www.bbc.com/news/technology-54146718

Chaudhri, A. B. (2003). *Web, Web-services, and database systems: NODe 2002 Web- and database-related workshops, Erfurt, Germany, October 7-10, 2002 revised papers* (Vol. 2593). Springer.

Chen, M., Preston, & Teubner. (2010). Information Systems Strategy: Reconceptualization, Measurement, and Implications. *MIS Quarterly*, *34*(2), 233–259. https://doi.org/10.2307/20721426

Christensen, L. T., Kärreman, D., & Rasche, A. (2019). Bullshit and Organization Studies. *Organization Studies*, *40*(10), 1587–1600. https://doi.org/10.1177/0170840618820072

Cohen, G. A. (2002). Deeper into Bullshit. In S. Buss & L. Overton (Eds.), *The Contours of Agency*. The MIT Press. https://doi.org/10.7551/mitpress/2143.003.0015

Cook, G., & Jardim, E. (2019). *Clicking Clean Virginia: The Dirty Energy Powering Data Center Alley*. Greenpeace Inc. https://www.greenpeace.org/usa/reports/click-clean-virginia/

Cook, G., Lee, J., Tsai, T., Kong, A., Johnson, B., & Jardim, E. (2017). Clicking Clean: Who is winning the race to build a green Internet? http://www.clickclean.org/international/en/

Coroamă, V. (2021). *Investigating the Inconsistencies Among Energy and Energy Intensity Estimates of the Internet: Metrics and Harmonising Values* (tech. rep.). Bern. https://www.aramis.admin.ch/Default?DocumentID=67656

Council of the European Union. (2011). Council Implementing Regulation (EU) No 282/2011 of 15 March 2011 laying down implementing measures for Directive 2006/112/EC on the common system of value added tax. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32011R0282

Croasmun, J. T., & Ostrom, L. (2011). Using Likert-Type Scales in the Social Sciences. *Journal of Adult Education*, *40*, "Article 1", 19–22.

Dalvi-Esfahani, M., Alaedini, Z., Nilashi, M., Samad, S., Asadi, S., & Mohammadi, M. (2020). Students' green information technology behavior: Beliefs and personality traits. *Journal of Cleaner Production*, *257*. https://doi.org/10.1016/j.jclepro.2020.120406

Diener, F., & Špaček, M. (2021). Digital Transformation in Banking: A Managerial Perspective on Barriers to Change. *Sustainability*, *13*(4), 2032. https://doi.org/10.3390/su13042032

Dresing, T., & Pehl, T. (Eds.). (2015). *Praxisbuch Transkription: Regelsysteme, Software und praktische Anleitungen für qualitative ForscherInnen* (6. Aufl.). Dr. Dresing und Pehl GmbH.

Ebrahimi, K., Jones, G. F., & Fleischer, A. S. (2014). A review of data center cooling technology, operating conditions and the corresponding low-grade waste heat recovery opportunities. *Renewable and Sustainable Energy Reviews*, *31*, 622–638. https://doi.org/10.1016/j.rser.2013.12.007

Ecosia. (2022). Ecosia is the search engine that plants trees. https://info.ecosia.org/what

Ericsson. (2020). A quick guide to your digital carbon footprint. https://www.ericsson.com/en/reports-and-papers/industrylab/reports/a-quick-guide-to-your-digital-carbon-footprint

Facebook. (2020). Sustainability Report 2020. https://sustainability.fb.com/report/2020-sustainability-report/

Frankfurt, H. G. (2005). *On Bullshit*. Princeton University Press. http://www.loc.gov/catdir/enhancements/fy0654/2004058963-d.html

Frankfurt, H. G. (2006). *On Truth*. Vintage Digital.

Frels, R. K., & Onwuegbuzie, A. J. (2013). Administering Quantitative Instruments With Qualitative Interviews: A Mixed Research Approach. *Journal of Counseling & Development*, *91*(2), 184–194. https://doi.org/10.1002/j.1556-6676.2013.00085.x

Garren, S. J., & Brinkmann, R. (2018). Sustainability Definitions, Historical Context, and Frameworks. In R. Brinkmann & S. J. Garren (Eds.), *The Palgrave Handbook of Sustainability* (pp. 1–18). Springer International Publishing. https://doi.org/10.1007/978-3-319-71389-2_1

Gilks, H., & Watson, E. (2019). Science Based Targets vs. Net Zero: what's the difference? Why do they matter? https://gresb.com/nl-en/science-based-targets-vs-net-zero-whats-the-difference-why-do-they-matter/

Gioia, D. A., Corley, K. G., & Hamilton, A. L. (2013). Seeking Qualitative Rigor in Inductive Research. *Organizational Research Methods*, *16*(1), 15–31. https://doi.org/10.1177/1094428112452151

Google. (2020). Realizing a carbon-free future: Google's Third Decade of Climate Action. Retrieved March 22, 2022, from https://www.gstatic.com/gumdrop/sustainability/carbon-free-by-2030.pdf

Google. (2021a). Environmental Report 2021. Retrieved April 25, 2022, from https://sustainability.google/reports/

Google. (2021b). Giving you more sustainable choices with Google. Retrieved March 28, 2022, from https://blog.google/outreach-initiatives/sustainability/sustainability-2021/

Grimm, D., Erek, K., & Zarnekow, R. (2013). Carbon Footprint of IT-Services - A comparative Study of energy consumption for Offline and Online Storage Usage.

Hacker, J., & Bodendorf, F. (2017). Conceptualising and Understanding User Behaviour in Enterprise Social Networks: A Qualitative Analysis. *ACIS 2017 Proceedings*, Article 41. https://aisel.aisnet.org/acis2017/41

Harvey, F. (2020). UN secretary general urges all countries to declare climate emergencies. *The Guardian*. https://www.theguardian.com/environment/2020/dec/12/un-secretary-general-all-countries-declare-climate-emergencies-antonio-guterres-climate-ambition-summit

Hernandez, A. A. (2019). An Empirical Investigation on the Awareness and Practices of Higher Education Students in Green Information Technology. *International Journal of Social Ecology and Sustainable Development*, *10*(2), 1–13. https://doi.org/10.4018/IJSESD.2019040101

Hintemann, R., & Hinterholzer, S. (2019). Energy Consumption of Data Centers Worldwide – How will the Internet become Green? *ICT4S*.

Hintemann, R., & Hinterholzer, S. (2020). *Videostreaming: Energiebedarf und CO2-Emissionen* (tech. rep.). Borderstep Institut für Innovation und Nachhaltigkeit. Berlin. https://www.borderstep.de/wp-content/uploads/2020/06/Videostreaming_2020.pdf

Hofmann, J., Ricci, C., & Ansu-Holz, D. (2021). IT und Nachhaltigkeit – eine Einführung. *HMD Praxis Der Wirtschaftsinformatik*, *58*(1), 6–23. https://doi.org/10.1365/s40702-020-00692-x

Hull, R., Benedikt, M., Christophides, V., & Su, J. (2003). E-services: A Look Behind the Curtain. *ACM Conferences, Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, 1–14. https://doi.org/10.1145/773153.773154

Idowu, S. O., Capaldi, N., Zu, L., & Gupta, A. D. (2013). *Encyclopedia of Corporate Social Responsibility*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-28036-8

Ijab, M. T., Molla, A., Kassahun, A. E., & Teoh, S. Y. (2010). Seeking the Green in Green IS: A Spirit, Practice and Impact Perspective. *PACIS 2010 Proceedings*, Article 46. http://aisel.aisnet.org/pacis2010/46

ITU. (2020). *Greenhouse gas emissions trajectories for the information and communication technology sector compatible with the UNFCCC Paris Agreement* (tech. rep.). Geneva, Switzerland. Retrieved December 2, 2021, from https://handle.itu.int/11.1002/1000/14084

ITU. (2021). *Measuring digital development: Facts and figures* (tech. rep.). Geneva, Switzerland. Retrieved May 5, 2022, from https://www

.itu.int/en/ITU-D/Statistics/Documents/facts/FactsFigures2021.pdf

Iyengar, M., David, M., Parida, P., Kamath, V., Kochuparambil, B., Graybill, D., Schultz, M., Gaynes, M., Simons, R., Schmidt, R., & Chainer, T. (2012). Server liquid cooling with chiller-less data center design to enable significant energy savings. *2012 28th Annual IEEE Semiconductor Thermal Measurement and Management Symposium (SEMI-THERM)*, 212–223. https://doi.org/10.1109/STHERM.2012.6188851

Jevons, W. S. (1866). *The Coal Question; an Inquiry Concerning the Progress of the Nation and the Probable Exhaustion of Our Coal-Mines*.

Jones, E. (2019). Rethinking Greenwashing: Corporate Discourse, Unethical Practice, and the Unmet Potential of Ethical Consumerism. *Sociological Perspectives*, *62*(5), 728–754. https://doi.org/10.1177/0731121419849095

Jong, M. D. T. d., Huluba, G., & Beldad, A. D. (2020). Different Shades of Greenwashing: Consumers' Reactions to Environmental Lies, Half-Lies, and Organizations Taking Credit for Following Legal Obligations. *Journal of Business and Technical Communication*, *34*(1), 38–76. https://doi.org/10.1177/1050651919874105

Jost, P., & Divitini, M. (2020). Game elicitation: exploring assistance in delayed-effect supply chain decision making. *ACM Digital Library. Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*, 1–10. https://doi.org/10.1145/3419249.3420154

Kamiya, G. (2020). The Carbon Footprint of Streaming Video: Fact-Checking the Headlines. *IEA Commentaries*. https://www.iea.org/commentaries/the-carbon-footprint-of-streaming-video-fact-checking-the-headlines

Kamiya, G. (2021). *Data Centres and Data Transmission Networks* (tech. rep.). IEA. Paris. https://www.iea.org/reports/data-centres-and-data-transmission-networks

Kaufman, M. (2020). The Devious Fossil Fuel Propaganda We All Use: A 'Successful, Deceptive' PR Campaign. *Mashable*. https://mashable.com/feature/carbon-footprint-pr-campaign-sham

Kern, E., Dick, M., Johann, T., & Naumann, S. (2011). Green Software and Green IT: An End Users Perspective. In P. Golinska, M. Fertsch, & J. C. Marx Gómez (Eds.), *Information Technologies in Environmental Engineering: New Trends and Challenges* (pp. 199–211, Vol. 3). Springer. https://doi.org/10.1007/978-3-642-19536-5_16

Koo, C., Chung, N., & Nam, K. (2015). Assessing the Impact of Intrinsic and Extrinsic Motivators on Smart Green IT Device Use: Reference Group Perspectives. *International Journal of Information Management*, *35*(1), 64–79. https://doi.org/10.1016/j.ijinfomgt.2014.10.001

Koomey, J., Berard, S., Sanchez, M., & Wong, H. (2011). Implications of Historical Trends in the Electrical Efficiency of Computing. *IEEE Annals of the History of Computing*, *33*(3), 46–54. https://doi.org/10.1109/MAHC.2010.28

Kuckartz, U. (2018). *Qualitative Inhaltsanalyse: Methoden, Praxis, Computerunterstützung* (4. Auflage). Beltz Juventa.

Kurkoon, P., Pimchangthong, D., & Boonjing, V. (2016). Environmental Awareness in Information Technology Adoption and Consumer Intention to Support Green Businesses: Research Agenda for Empirical Study. *Information Systems Development: Complexity in Information Systems Development (ISD2016 Proceedings)*.

Kurkoon, P., Pimchangthong, D., & Boonjing, V. (2018). Green Support and the Use of Information Technology Products. *Designing Digitalization (ISD2018 Proceedings)*. http://aisel.aisnet.org/isd2014/proceedings2018/Transforming/2

Kutani, A. I., Hennicke, P., Gericke, N., & Bunge, F. (2021). *Energy and Climate Policy in the Post COVID-19 Era Comparative Analyses on Germany and Japan* (tech. rep.). Retrieved December 2, 2021, from http://www.gjetc.org/wp-content/uploads/2021/04/GJETC-Study_Energy-policy-Post-Covid_FINAL-210428.pdf

Kvasnicova, T., Kremenova, I., & Fabus, J. (2016). From an Analysis of e-services Definitions and Classifications to the Proposal of New e-service Classification. *Procedia Economics and Finance*, *39*, 192–196. https://doi.org/10.1016/S2212-5671(16)30282-9

Lange, S., Pohl, J., & Santarius, T. (2020). Digitalization and Energy Consumption. Does ICT Reduce Energy Demand? *Ecological Economics*, *176*, 106760. https://doi.org/10.1016/j.ecolecon.2020.106760

Leung, N. K. Y., Lau, S. K., Shamsub, H., & Lau, S. Y. (2018). A Study of Perception Factors that Affect Green IT Behavior. *AMCIS 2018 Proceedings*, Article 1. https://aisel.aisnet.org/amcis2018/GreenIS/Presentations/1

Likert, R. (1932). A Technique for the Measurement of Attitudes. *Archives of Psychology*, *22*, "Article 140", 5–55.

Loeser, F. (2013). Green IT and Green IS: Definition of Constructs and Overview of Current Practices. *AMCIS 2013 Proceedings*, Article 4. https://aisel.aisnet.org/amcis2013/GreenIS/GeneralPresentations/4

Luks, F. (2017). The Ugly, the Bad, and the Good: Bullshit as Discourse, Accursed Share, and Lubricant. *Journal of Extreme Anthropology*, *1*(1), 85–91. https://doi.org/10.5617/jea.5378

Lyon, T. P., & Maxwell, J. W. (2011). Greenwash: Corporate Environmental Disclosure Under Threat of Audit. *Journal of Economics & Management Strategy*, *20*(1), 3–41. https://doi.org/10.1111/j.1530-9134.2010.00282.x

Lyon, T. P., & Montgomery, A. W. (2015). The Means and End of Greenwash. *Organization & Environment*, *28*(2), 223–249. https://doi.org/10.1177/1086026615575332

Malmodin, J., & Lundén, D. (2018). The Energy and Carbon Footprint of the Global ICT and E&M Sectors 2010–2015. *Sustainability*, *10*(9), 3027. https://doi.org/10.3390/su10093027

Markle, G. L. (2013). Pro-Environmental Behavior: Does It Matter How It's Measured? Development and Validation of the Pro-Environmental Behavior Scale (PEBS). *Human Ecology*, *41*(6), 905–914. https://doi.org/10.1007/s10745-013-9614-8

Marks, L. U., Clark, J., Livingston, J., Oleksijczuk, D., & Hilderbrand, L. (2020). Streaming Media's Environmental Impact [Advance online publication]. *Media+Environment*. https://doi.org/10.1525/001c.17242

Masanet, E., Shehabi, A., Lei, N., Smith, S., & Koomey, J. (2020). Recalibrating Global Data Center Energy-Use Estimates. *Science*, *367*(6481), 984–986. https://doi.org/10.1126/science.aba3758

Maxime, E.-H., & Jean-Noël, G. (2020). Did The Shift Project really overestimate the carbon footprint of online video? Our analysis of the IEA and CarbonBrief articles. https://theshiftproject.org/wp-content/uploads/2020/06/2020-06_Did-TSP-overestimate-the-carbon-footprint-of-online-video_EN.pdf

Mayring, P., & Fenzl, T. (2019). Qualitative Inhaltsanalyse. In N. Baur & J. Blasius (Eds.), *Handbuch Methoden der Empirischen Sozialforschung* (pp. 633–648). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-21308-4_42

Meszaros, J., & Buchalcevova, A. (2017). Introducing OSSF: A Framework for Online Service Cybersecurity Risk Management. *Computers & Security*, *65*, 300–313. https://doi.org/10.1016/j.cose.2016.12.008

Molla, A., Abareshi, A., & Cooper, V. (2014). Green IT beliefs and pro-environmental IT practices among IT professionals. *Information Technology & People*, *27*(2), 129–154. https://doi.org/10.1108/ITP-10-2012-0109

Molla, A., Pittayachawan, S., & Cooper, V. A. (2009). IT and Eco-sustainability: Developing and Validating a Green IT Readiness Model. *ICIS 2009 Proceedings*, 141. http://aisel.aisnet.org/icis2009/141

Moore, G. E. (1965). Cramming more components onto integrated circuits [Article 8]. *Electronics Magazine*, (38), 114.

Moore, G. E. (1975). Progress in Digital Integrated Electronics. *IEDM Tech. Digest*, 11–13.

Moosbrugger, H., & Kelava, A. (Eds.). (2020). *Lehrbuch. Testtheorie und Fragebogenkonstruktion* (3., vollständig neu bearbeitete, erweiterte und aktualisierte Auflage). Springer.

Nash, K., & Wakefield, R. L. (2021). The Role of Identity in Green IT Attitude and Intention. *Journal of Computer Information Systems*, 1–11. https://doi.org/10.1080/08874417.2021.1960221

Netflix. (2021). Environmental Social Governance Report 2021. Retrieved April 25, 2022, from https://about.netflix.com/en/sustainability

Oates, B. J. (2006). *Researching Information Systems and Computing*. Sage Publications Ltd.

Obringer, R., Rachunok, B., Maia-Silva, D., Arbabzadeh, M., Nateghi, R., & Madani, K. (2021). The Overlooked Environmental Footprint of Increasing Internet Use. *Resources, Conservation and Recycling*, *167*, 105389. https://doi.org/10.1016/j.resconrec.2020.105389

Orange, E., & Cohen, M. A. (2010). From Eco-Friendly to Eco-Intelligent. *The Futurist*, *44*(5), 28–32.

Pakkala, D., & Spohrer, J. (2019). Digital Service: Technological Agency in Service Systems. In T. X. Bui (Ed.), *Proceedings of the 52nd Annual Hawaii International Conference on System Sciences* (pp. 1886–1895). https://doi.org/10.24251/HICSS.2019.229

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2015). On the Reception and Detection of Pseudo-Profound Bullshit. *Judgment and Decision Making*, *10*(6), 549–575. https://econpapers.repec.org/RePEc:jdm:journl:v:10:y:2015:i:6:p:549-563

Petrocelli, J. V. (2018). Antecedents of Bullshitting. *Journal of Experimental Social Psychology*, *76*, 249–258. https://doi.org/10.1016/j.jesp.2018.03.004

Purvis, B., Mao, Y., & Robinson, D. (2019). Three Pillars of Sustainability: In Search of Conceptual Origins. *Sustainability Science*, *14*(3), 681–695. https://doi.org/10.1007/s11625-018-0627-5

Robinson, M., & Crowther, D. (2001). Environmental Science Literacy in Science Education, Biology & Chemistry Majors. *The American Biology Teacher*, *63*(1), 9–14. https://doi.org/10.2307/4451023

Rowley, J. (2012). Conducting Research Interviews. *Management Research Review*, *35*(3/4), 260–271. https://doi.org/10.1108/01409171211210154

Rutsch, J. (2020). Greenwashing oder Vorbild? Die Klimaziele von Google, Facebook, Amazon und Co. *hr-iNFO*. https://www.ardaudiothek.de/episode/hr-info-wissenswert/greenwashing-oder-vorbild-die-klimaziele-von-google-facebook-amazon-und-co/hr-info/86561084/

Sands, H., & Morison, B. (2020). Greenwashing in the Information Industry. *The IJournal: Graduate Student Journal of the Faculty of Information*, *5*(2). https://doi.org/10.33137/ijournal.v5i2.34413

Sandvine. (2020). The Global Internet Phenomena Report COVID-19 Spotlight. Retrieved December 2, 2021, from https://www.sandvine.com/hubfs/Sandvine_Redesign_2019/Downloads/2020/Phenomena/COVID%20Internet%20Phenomena%20Report%20200200507.pdf

Santarius, T., Pohl, J., & Lange, S. (2020). Digitalization and the Decoupling Debate: Can ICT Help to Reduce Environmental Impacts While the Economy Keeps Growing? *Sustainability*, *12*(18), 7496. https://doi.org/10.3390/su12187496

SBTi. (2020). *SBTi criteria and recommendations*. Retrieved December 16, 2021, from https://sciencebasedtargets.org/resources/legacy/2019/03/SBTi-criteria.pdf

SBTi. (2021). *Taking the Temperature: Assessing and Scaling-Up Climate Ambition in the G7 Business Sector*.

Schiettekatte, A. F. R. (2021). *Environmental Impact of Producing and Operating Server Infrastructure: A Systematic Literature Review* [Bachelor Degree Project]. University of Skövde [[Bachelor Degree Project in Information Technology]].

Schmidt, N.-H. (2011). *Environmentally Sustainable Information Management: Theories and Concepts for Sustainability, Green IS, and Green IT* (1. Aufl., Vol. 64). Cuvillier. https://cuvillier.de/de/shop/publications/295-environmentally-sustainable-information-management

Schreier, M. (2014). Varianten qualitativer Inhaltsanalyse: Ein Wegweiser im Dickicht der Begrifflichkeiten. *Qualitative Social Research*, *15*(1), Article 18. https://doi.org/10.17169/fqs-15.1.2043

Seele, P., & Gatti, L. (2017). Greenwashing Revisited: In Search of a Typology and Accusation-Based Definition Incorporating Legitimacy Strategies. *Business Strategy and the Environment*, *26*(2), 239–252. https://doi.org/10.1002/bse.1912

Spicer, A. (2018). *Business Bullshit*. Routledge.

Spotify. (2021). Equity and Impact Report 2021. Retrieved April 25, 2022, from https://www.lifeatspotify.com/diversity-equity-impact/climate-action

Stevenson, H. (2021). Reforming Global Climate Governance in an Age of Bullshit. *Globalizations*, *18*(1), 86–102. https://doi.org/10.1080/14747731.2020.1774315

TeleGeography. (2021). Global Internet Geography: Executive Summary. Retrieved December 2, 2021, from https://www2.telegeography.com/hubfs/assets/product-tear-sheets/product-page-content-samples/global-internet-geography/telegeography-global-internet-geography-executive-summary.pdf

TerraChoice. (2009). The Seven Sins of Greenwashing: Environmental Claims in Consumer Markets.

The Shift Project. (2019). Climate Crisis: The Unsustainable Use of Online Video. Retrieved March 21, 2022, from https://theshiftproject.org/en/article/unsustainable-use-online-video/

The Shift Project Materials. (2019). 1byte Model Applied to Video Traffic [[video+ materials]]. Retrieved May 2, 2022, from https://theshiftproject.org/wp-content/uploads/2020/06/Video-Materials-1byte-Model-applied-to-video-traffic-updated-2020.xlsx

U.S. EPA. (2021). Clean Energy Finance: Using Renewable Energy Certificates to Achieve Local Environmental Goals. https://www.epa.gov/sites/default/files/2021-04/documents/usepa_renewableenergycertificates_april_2021.pdf

Vrana, J., & Singh, R. (2021). Digitization, Digitalization, and Digital Transformation. In N. Meyendorf, N. Ida, R. Singh, & J. Vrana (Eds.), *Handbook of Nondestructive Evaluation 4.0* (pp. 1–17). Springer International Publishing. https://doi.org/10.1007/978-3-030-48200-8_39-1

Wagner, K., & Schramm-Klein, H. (2019). Alexa, Are You Human? Investigating Anthropomorphism of Digital Voice Assistants – A Qualitative Approach. *ICIS 2019 Proceedings*, Article 7. https://aisel.aisnet.org/icis2019/human_computer_interact/human_computer_interact/7

Warde, P. (2011). The Invention of Sustainability. *Modern Intellectual History*, *8*(1), 153–170. https://doi.org/10.1017/S1479244311000096

Widdicks, K., Hazas, M., Bates, O., & Friday, A. (2019). Streaming, Multi-Screens and YouTube. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3290605.3300696

Junior Management Science

www.jums.academy
ISSN: 2942-1861



Board Gender Diversity: Evidence From Indonesia

Nadhilla Mazaya

Humboldt University of Berlin

Abstract

Board gender diversity continues to gain global attention, alongside a growing percentage of female board members in public companies. While board gender quotas have played a role in this increase, countries without such mandates have also experienced similar growth. This raises an important question: *What drives companies to appoint women to corporate boards in the absence of compulsory regulations?* Primarily, this paper examines the relationship between foreign institutional investors and female board members in supervisory and management boards within Indonesian public firms. This study analyzes data from 147 companies between 2019 and 2022 using OLS regression with lead and control variables. In contrast to the belief, the findings show that foreign institutional investors have a relatively low to no influence in shaping board gender diversity on each board. This lack of influence suggests that other factors may significantly affect companies' decisions to hire women on boards, highlighting the necessity to investigate these additional factors.

Keywords: board gender diversity; corporate boards; foreign investors

1. Introduction

In the last two decades, there has been a notable increase in the representation of women on corporate boards across the world. Since the introduction of the female board quota in Norway in 2003, many countries have followed suit by setting mandatory quotas, driving the growth in the percentage of women on boards (Matanda et al., 2023). Moreover, this increase could also be associated with the general discussions on the benefits of adding women onto boards, which have potentially attracted companies to increase the proportion of female board members. Previous studies have discussed that the presence of women on boards helps to improve the governance of companies, leading to better firm performance (Adams & Ferreira, 2009; Adams & Funk, 2012). Nevertheless, this increasing trend is not limited to countries with mandatory regulations. Countries with either voluntary or no mandates on board gender diversity have similarly witnessed a rise in female board members (Matanda et al., 2023), suggesting the possibility of other underlying factors that influence such growth beyond mandatory regulations. Arguably, this trend raises an intriguing inquiry into the mechanisms behind such increases. Therefore, in this study, I explore the underlying drivers influencing board gender diversity.

Specifically, I raise the question of what drives companies to appoint women to corporate boards in the absence of compulsory regulations. The answers to this question deepen the understanding of the factors contributing to the growing rate of female board members. This knowledge is valuable, particularly for stakeholders, informing them of plausible channels to enhance board gender diversity, especially in companies where female representation is notably low. Further, the answers provide deeper insights to encourage policymakers to support these drivers, potentially through more directive legislation.

Albeit limited, previous research has explored the determinants of board gender diversity, both in environments with and without formal mandates. A study by Kirsch (2018) re-

I want to express my gratitude to God the Almighty for His blessings and guidance, which helped me complete my master's thesis. Further, I specifically extend my gratitude to Prof. Dr. Ulf Brüggemann and Lisa Lüttke, M.Sc., for their support and for giving me valuable and meticulous suggestions during and after the time of writing. I am also deeply grateful to my family and friends for their presence and continuous support.

DOI: https://doi.org/10.5282/jums/v10i3pp609-630

© The Author(s) 2025. Published by Junior Management Science.
This is an Open Access article distributed under the terms of the CC-BY-4.0 (Attribution 4.0 International). Open Access funding provided by ZBW.

veals that macro-level country characteristics such as high gender equality and access to work-family balance, as well as pressure from shareholders and formal institutional bodies, contribute to a high level of board gender diversity. Further, this study indicates that industry characteristics, firm performance, and social connections are associated with the presence of female board members. Another study suggests that laws and the cultural and religious beliefs of the society tend to be correlated with board gender diversity in many Asian countries (Shams, 2019). In addition, research conducted by Fauver et al. (2022) discusses the association between a high representation of female board members and the presence of foreign institutional shareholders among public companies in various countries, particularly after the enforcement of board gender diversity reforms.

Shifting from multi-setting research to a more specific setting, Abdullah (2014) highlights the positive link between board size and a higher proportion of female directors among Malaysian companies listed in 2007. However, this study also suggests that companies with poor financial performance tend to have more female directors than those with strong financial performance. Furthermore, a previous study by Rastad and Dobson (2022) highlights the effectiveness of shareholder pressure in increasing the percentage of female directors among US public companies.

Although prior studies have explored various drivers, they have neither engaged in nor adequately discussed the drivers of board gender diversity in Indonesia, a country known for having a vast female labor force but a low representation of women on boards (Deloitte Global, 2022). Over the last decade, the country reports a proportion of female directors between 3% and 12% (Matanda et al., 2023) and a proportion of female commissioners between 11% and 18% (Darmadi, 2011; Fauziah et al., 2022).[1] Interestingly, despite the low proportions, this literature reveals that Indonesia has experienced a growing trend in the percentage of female board members. This insight offers a suitable setting to address the research question introduced earlier, considering the absence of formal mandates on board gender diversity in Indonesia.

Further, there appears to be an absence of previous studies discussing the association between foreign institutional investors and the presence of women on boards in the Indonesian context. Shareholders possess the power to shape board gender diversity through their voting rights when electing board members. Additionally, prior literature indicates that shareholders successfully push companies to increase the female proportion on boards through various channels such as private negotiations, proposals, and public media. However, in many cases, there is a time lag between the initial requests to increase board gender diversity and the actual implementation (Bauer et al., 2015; Gormley et al., 2023; Rastad & Dobson, 2022). Further, previous studies also emphasize that, among other types of investors, foreign institutional investors are more likely to influence companies to increase female representation on boards, especially those from countries with higher levels of board gender diversity (Fauver et al., 2022; Tang & Zhang, 2021).

To further investigate this association, I employ OLS and Linear Probability Regression models. After conducting the analyses, I find that foreign institutional investors are not associated with the presence of female commissioners on boards but are associated with the presence of female directors. This implies that companies appoint at least one female director potentially due to the presence of foreign institutional investors. However, this finding no longer holds after controlling for within-firm and over-time variations. Furthermore, I find that board and firm size contribute to the level of board gender diversity among Indonesian companies. Additionally, my data highlight the consistently low proportion of female board members in Indonesian public companies.

As implied in the earlier discussion, my study contributes to the literature by focusing on Indonesia, a previously underexplored setting, and by extending the limited research on the drivers of board gender diversity in both global and Indonesian academic studies. Further, my study underscores the need for greater attention to increasing female representation on boards among Indonesian public companies, potentially through the initiation of internal strategies within firms and the enforcement of legislation by the authorities.

2. Conceptual Framework

2.1. Board Gender Diversity

Board gender diversity, arguably, emerges as a potential rationale for understanding what contributes to board effectiveness while simultaneously serving as a strategy to reduce the glass ceiling in the workplace (Adams, 2016). For years, there has been a prevailing belief that gender-diverse boards contribute significantly to improved governance, particularly in corporate decision-making processes. Previous studies reveal that women bring unique perspectives and approaches to both financial and non-financial aspects of business, enriching the pool of information available to the board (Adams & Ferreira, 2009; Adams & Funk, 2012). This wide range of information pool enhances alternatives for strategic decision-making, leading to better governance. In addition, the idea of increasing female representation on boards has also been circulating among policymakers, leading to the establishment of female board members quotas and public campaigns on board gender diversity.

The first legislation for board gender quota was introduced in 1999 in Israel, requiring state-owned companies to have at least one female director on board. However, it was not until the mid-2000s that this issue gained significant interest worldwide (Kirsch, 2018). In 2003, the government of Norway introduced legislation on female board quotas for public companies, which was enforced in 2008. After

[1] Indonesia adopts two-tier board system with a supervisory board known as the Board of Commissioners (the BOC) and an executive board known as the Board of Directors (the BOD). These boards are detailed further in the second chapter of this study.

this successful example, many countries, such as Germany, France, and Belgium, followed the same path and introduced their national regulations on female board members, particularly for listed companies (Fauver et al., 2022).

Within the scope of reforms and regulations, some countries enforce mandatory quotas for female board representation, while others prefer softer policies. Currently, around 17 countries have set compulsory quotas for public companies, aiming for female participation rates between 25% and 40% (Matanda et al., 2023). Within these quotas, some countries enacted penalties in cases of non-compliance. For instance, in Norway, a company is prohibited from registering if it fails to satisfy the mandatory 40% quota for female board members, and consistent non-compliance after several warnings may lead to the company being dissolved (Storvik & Teigen, 2010). Meanwhile, around 13 countries impose voluntary measures to enhance female representation on boards. These soft-law strategies may include mandatory disclosures about board gender composition, internal approaches to increase the number of women on boards, self-determined company quotas, or combinations of these strategies (Terjesen et al., 2015).

These reforms have led to a notable global increase in the proportion of female board members. Figure 1 reveals that from 2018 to 2022, countries with mandatory quotas have experienced a consistent increase in female board representation, with percentages rising from 21% in 2018 to almost 30% by 2022. In addition, this figure also illustrates a high level of commitment from companies in countries with softer regulations, as the trend and percentages closely resemble those in countries with mandatory quotas. Meanwhile, countries without reforms have a lower rate of female board participation despite having a similar increasing trend. Notably, these countries have approximately six to ten percent lower rates of female directors than the others over these five years. Essentially, the trend indicates that the role of board gender diversity reforms in the growth of female board members by serving as an additional external pressure.

While these legislative efforts have been largely effective, they are subject to ongoing debate. Critics have voiced concerns that mandatory quotas may limit an owner's right to select board members, potentially leading to biased choices (Fauver et al., 2022). Further, enforcing penalties for non-compliance is often associated with the success of these quotas (Storvik & Teigen, 2010). Additionally, in the case of non-mandatory regulations, the increase in the proportion of female board members could be linked to the desire of companies to maintain their public image (Abdullah, 2014; Perrault, 2015). Notably, as depicted in the figure note of Figure 1, the countries that have implemented these reforms are primarily developed, with fewer developing countries. This pattern suggests that the success of such reforms may depend on the unique institutional characteristics and labor market conditions that are more typical of developed nations.

Ultimately, board gender diversity is an essential aspect of corporate governance. Although some debates exist regarding this concept, the efficacy of female board members in enhancing company governance cannot be overlooked. Exploring the drivers behind board gender diversity and its importance for companies might be further underscored. As previously discussed, government reforms have been associated with an increase in the proportion of women on boards, yet the same upward trend is observed in countries without such regulations. This insight suggests the presence of additional factors driving this progress, which I will further explore in the subsequent sub-chapters.

2.2. Literature Review on Determinants of Board Gender Diversity

One previous study discusses the drivers of board gender diversity from several perspectives across the macro, meso, and micro levels (Kirsch, 2018). This research analyzed 310 articles, published in 135 journals worldwide between 1981 and 2016, on board gender diversity, including, but not limited to, mechanisms that drive board gender diversity. The study reveals that at the macro level, board gender diversity is likely to be driven by gender equality and regulations alleviating women's access to balance work and family commitments. Additionally, this study discusses the influence of key agents of change, including shareholders, executive search firms, stock exchanges, and the media, in disseminating board gender diversity. At the meso level, the study emphasizes industry characteristics and firm performance as drivers for appointing female members to the board, particularly when the effects of having women on the board benefit the firms. At the micro level, this study considers social networks and demographic characteristics, where the presence of women on the board tends to be linked to the presence of women in the nomination committees. However, given the scope of this literature review, which encompasses a wide variety of research settings, the conclusions drawn may be overly generalized and may not directly apply to one particular country's setting.

Research on the drivers of board gender diversity in a more specific setting is conducted by Shams (2019). This study examines the factors influencing female participation on boards in 13 different Asian countries in a single year from 2014 to 2016 based on the data availability of each country.[2] Utilizing quantitative and qualitative methods, this research analyzes the influence on female board participation in 5,054 firms, comprised of firms listed on major stock exchanges of each country. This study shows that in formal settings, the likelihood of using civil law rather than common law increases female board representation by 0.068, which might be attributed to labor welfare regulations facilitating work-family balance. Further, it highlights the positive effects of female education and literacy rates on board gender

[2] The data for Bhutan, India, Pakistan, Philippines, Singapore, and Vietnam are collected for the year 2015, while the data for Bangladesh, Indonesia, Malaysia, Sri Lanka, and Thailand are collected for the year 2014. Lastly, the data for Nepal and Maldives are collected for the year 2016. Additionally, this study uses Tobit regression and Fuzzy set Qualitative Comparative Analysis (fsQCA) to examine the factors shaping female representatives in the boardroom.

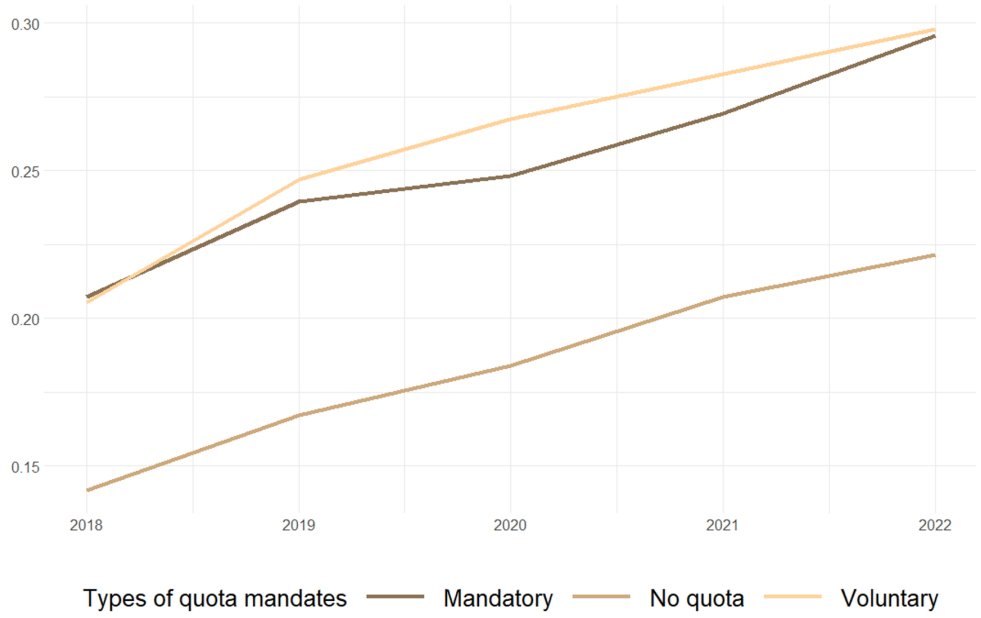


Figure 1: Proportion of Female Directors on Board Worldwide by Types of Quota Mandates

This graph categorizes countries into: (1) Mandatory quotas, legally binding board gender requirements, comprising 14 developed and 3 developing countries; (2) Voluntary quotas, non-binding guidelines or narrative disclosure on women's board/executive roles, comprising 12 developed and 1 developing countries; and (3) No quotas, no formal rules, comprising 7 developed and 9 developing countries. Source: Statistics from Deloitte Global (2022) and Matanda et al. (2023).

diversity, especially in Vietnam, the Philippines, and Thailand. In informal settings, the study reveals that the cultural and religious attitudes of the citizens play a role in driving female board representation. Countries viewing women as elite members of society tend to have higher female board members. Meanwhile, countries where more than 50% of the population practices a religion are more likely to have fewer female directors by 0.144 percentage points. Nevertheless, despite focusing on a more specific context, this study generalizes factors influencing board gender diversity without differentiating the drivers among each Asian country. Each country has its unique institutional features that might not be suitable for direct comparisons.

Abdullah (2014) studies factors shaping board gender compositions among large public firms in Malaysia. It highlights that among non-financial companies listed on the Malaysian stock exchange in 2007, firms are more likely to have at least one female director if the companies have a large board size. However, the study reports no observable association between the likelihood of selecting female directors and firm size, age, ethnic diversity, or leverage. Conversely, the study shows a negative association between board gender diversity and firm performance. Interestingly, the study reveals that appointing females to boards serves as a token gesture to comply with shareholders' interests in gender diversity. The data show that a significant number of female directors are connected to controlling shareholders or other board members and are prevalent in family-owned or government-owned firms. However, similar to the study conducted by Shams (2019), this study focuses only on a single period, which may overlook the timing of the factors in driving board gender diversity. Perhaps the significance and impact of these drivers might vary across different periods.

Rastad and Dobson (2022) conduct a longitudinal analysis, providing a more comprehensive understanding of the drivers of board gender diversity across different periods. They find that board gender diversity among US public companies is influenced by both internal and external pressures exerted by shareholders. They explain that internal pressure arises from private negotiations between the management and shareholders of the companies. In contrast, external pressure stems from proposals submitted by the shareholders to be voted on at the general meeting of shareholders. Upon studying diversity-related proposals between 1997 and 2013, they report that these pressures have significant effects, leading to a greater proportion of female board members. However, sometimes, these effects have a lag between the time of submission and implementation of the proposal. A similar finding is reported in a study conducted by Fauver et al. (2022), which highlights the role of foreign institutional investors in driving board gender diversity. Using a difference-in-differences design on companies from 43 countries between 2000 and 2017, they find that foreign institutional investors (lag) are associated with the increase in the proportion of female directorships, especially in countries with high social equity norms. However, these increases are becoming prominent only after the enforcement of board gender diversity reforms, emphasizing the significance of formal regulations in board gender diversity.

Among the potential drivers of board gender diversity discussed earlier, I specifically focus on the influence of shareholders in appointing women to boards. In the absence of mandatory regulations, companies might still face pressure to improve board gender diversity by non-enforceable mechanisms such as international expectations or social norms. While social norms offer valuable insights into shaping board gender diversity, researching them might be time-consuming and arduous.[3] In contrast, analyzing internal and external shareholder pressure provides a more structured and measurable understanding of how companies respond to diversity demands. Shareholders have the power to shape the gender composition of company boards through their voting rights. Although voting mechanisms and regulations may vary across countries, shareholders in most jurisdictions are capable of electing board members from a pool of candidates during the general meeting of shareholders (ISS, n.d.). While previous studies have provided an essential foundation for understanding the dynamics between shareholder actions and the appointment of women to boards, they also highlight the need for further exploration, particularly in different settings and time frames. Consequently, I facilitate this calling and explore the association between shareholders and board gender diversity in a different setting and different time frames, as detailed in the subsequent sections.

2.3. Underlying Theory: Shareholder Activism

Shareholders possess the power to influence the gender composition of corporate boards not only through their voting rights but also via the mechanism of shareholder activism. The framework of shareholder activism is derived from the voice vs. exit strategy by Albert O. Hirschman (Goodman et al., 2014). Hirschman defines "voice" as a political response to address inefficiencies in a company's performance, while "exit" represents an economic response where dissatisfaction leads shareholders to sell their shares. In shareholder activism, "voice" is the strategy shareholders use to drive change when they see potential for improvement. Otherwise, they may opt to exit by divesting their shares.

Various methods exist for activist shareholders, i.e., those who express their opinions to management, to employ the voice strategy. Engaging in dialogue with management is often viewed as the first strategic step in initiating change (Ryan & Schneider, 2002). This approach enables activist shareholders and management to reach a mutual understanding, taking into account the needs and interests of both parties. While this method is relatively straightforward, it might not always yield effective results, prompting shareholders to seek other methods.

The second preferred method is submitting shareholder proposals during the general meeting of shareholders. These proposals outline activist shareholders' concerns regarding the company's financial and non-financial performance. Typically, they include discussions of the specific issues, their importance and rationale, timelines for implementation, and how these concerns align with shareholders' broader interests, considering the company's resources (PRI, 2023). These proposals are either voted on during the meetings or withdrawn beforehand through mutual agreement between the proposers and management.

Moving beyond direct interactions, public confrontation and the strategic use of media could also be powerful tools for advocating change. A study revealed an interview with a board member suggesting that companies are highly responsive to news and social media, particularly on sensitive issues (Perrault, 2015). However, using public confrontation might put activist shareholders and companies in a difficult position, especially when using negative connotations. Therefore, this method is typically preferred if the two previous methods fail (Ryan & Schneider, 2002).

Last but not least, if the voice strategy proves unsuccessful, activist shareholders could adopt two additional methods. According to Admati and Pfleiderer (2009), shareholders might either sell their shares, a strategy known as the "Wall Street Walk," or threaten to sell their shares. However, the first method often requires solid insider information, which can be challenging to obtain and verify. Therefore, the latter method is generally more advantageous, especially when divestments are costly. Large shareholders frequently employ this type of threat to pressure companies into addressing their concerns and enacting changes.

2.4. Shareholders on Board Gender Diversity

To enhance the monitoring of top executives and reduce agency costs, shareholders appoint board members who can best protect their interests. Increasing gender diversity on traditionally male-dominated corporate boards may strengthen shareholders' influence over the board. The presence of women directors often fosters a culture of critical inquiry, reducing the likelihood of management decisions being approved without thorough scrutiny (Selby, 2000). Moreover, including women in the board's talent pool can contribute to better information flow and innovation, thereby improving the board's strategic decision-making (Adams & Ferreira, 2009; Adams & Funk, 2012).

Public opinion and media attention play a role in driving shareholders' interest in board gender diversity, with frequent press coverage highlighting its benefits. It is not surprising that shareholders respond positively to companies with gender-diverse boards, as their reactions reflect the broader public opinion that emphasizes the value of gender diversity in corporate governance (Kang et al., 2010). For instance, The Business Times (Singapore) published an article highlighting the advantages of men and women collaborating in the boardroom to enhance organizational performance (Quek, 2024).

Shareholders often employ the voice strategy to improve gender composition on boards. Initially, they engage in private negotiations with companies, requesting a more gender-

[3] Researching social norms might require, at the minimum, doing surveys or field studies. While such research could provide detailed insights and uncover interesting findings, it could also present challenges, particularly in terms of geographic scope and data collection.

diverse board (Perrault, 2015). Should these efforts not yield the desired outcomes, they file board gender diversity proposals with requests such as a clear diversity policy and implementation timeline (Rastad & Dobson, 2022). These proposals are later submitted at general shareholder meetings for voting or withdrawn beforehand. Most of them, however, are either voted down or withdrawn before the meeting (Bauer et al., 2015; Rojas et al., 2009).

While these proposals rarely receive immediate favorable results, studies show that rejections or withdrawals don't necessarily indicate failed attempts to increase the proportion of women on boards (Bauer et al., 2015; Rastad & Dobson, 2022).[4] They have a lag effect, where the implementation takes place around a year after the submission. Alternatively, shareholders initiate public campaigns to disseminate awareness of board gender diversity and increase female representation on corporate boards.[5]

However, the level of commitment to improving board gender diversity varies, particularly between foreign and local shareholders, and it depends on the country of origin. Board gender diversity tends to resonate more in developed countries, where there is greater awareness of gender equality.[6] This difference in commitment often leads foreign shareholders from developed countries to be more proactive in advocating for gender diversity, bringing standards closer to those in their home countries.

In response, firms are keen to align with the interests of these foreign shareholders, especially when they play a significant role in the company's ownership structure. Aligning with these expectations not only helps companies enhance their reputation and performance in international markets but also enables them to adopt more sophisticated governance practices. For instance, embracing board gender diversity can be seen as a strategic move to meet the global standards that foreign shareholders advocate, thereby increasing the firm's position in the worldwide market.

The proactive efforts of foreign shareholders often lead to tangible improvements in board gender diversity. A study analyzing board gender diversity across 25 countries, including both developed and developing nations, found that foreign shareholders from countries with higher gender equality scores exert more pressure on companies in those countries to increase the proportion of female board members (Fauver et al., 2022). Similarly, research on Chinese firms shows that the growing presence of foreign affiliates has been associated with a significant rise in the number of women in managerial roles. This trend underscores the broader impact of foreign involvement—whether through shareholders or affiliates—on promoting gender diversity within corporate leadership (Tang & Zhang, 2021).

On the other hand, while possessing a richer set of information about local companies, domestic shareholders might be more focused on short-term financial returns or traditional practices. Their familiarity with local norms might sometimes result in a limited perspective regarding gender diversity, which they might perceive as less critical within the business environment. A previous study highlights that, even when facing pressure from board gender quotas, domestic shareholders tend to remain indifferent to board gender diversity and have limited influence in fostering a more gender-diverse board (Fauver et al., 2022). However, a significant number of foreign shareholders might drive a shift in this stance. As mentioned earlier, foreign shareholders often bring a broader perspective and higher gender equality standards, which could prompt domestic shareholders to update their beliefs.

Existing literature supports this dynamic, indicating that under the framework of cultural spillover, local firms are likely to adjust their views and practices when exposed to differing perspectives from foreign shareholders (Tang & Zhang, 2021).[7] Nevertheless, foreign shareholders still need the local expertise of domestic shareholders to grasp the unique aspects of the local business context. By combining local insights with their broader perspectives, companies might create a more comprehensive strategy for addressing gender diversity issues and handle the challenges of doing business in that specific location more effectively.

Foreign shareholders from developed countries, particularly those from the US, are accustomed to shareholder activism and often submit proposals to advocate for more female representation in the boardroom.[8] However, while this trend is prevalent, the impact of such activism can differ significantly depending on the type of shareholder involved. For

[4] Withdrawn proposals, in particular, often imply that the management accept the requests of the proposing shareholders (Bauer et al., 2015). Additionally, Rastad and Dobson (2022) highlight that the impact on board gender diversity is stronger for withdrawn proposals among public companies in the US. Around 40% of these proposals are implemented within a year, while approximately 25% of rejected proposals are put into practice in the following year.

[5] One notable example is the "Fearless Girl" campaign advocated by State Street Global Advisors in March 2017. Alongside other asset managers like Blackrock and Vanguard, this campaign not only increased the percentage of female board members in US public companies but also prompted companies to appoint their first female directors in instances where none had served previously (Gormley et al., 2023).

[6] As shown in Figure 1, developed countries such as Australia, Denmark, Germany, Norway, and Singapore tend to have a higher proportion of female board members compared to developing countries such as Brazil, China, Indonesia, Poland, and Turkey. Developed countries tend to have a higher awareness of gender equality, which leads to the establishment of female board mandates among these countries.

[7] They argue that the presence of foreign-invested enterprises (FIEs) in the same market as local firms might gradually influence domestic firms' beliefs by creating competitive pressures and offering new strategies to imitate. This competitive advantage includes both the cost pressures FIEs introduce and the potential for domestic firms to learn and adopt more effective practices. For instance, if many FIEs hold less biased views about female workers, domestic firms might reduce their biases and employ more women. However, the overall impact on domestic firms depends on the nature of the FIEs' beliefs and the number of FIEs in the market.

[8] The three largest asset management firms publicly disclose their voting policies aimed at increasing female representation on corporate boards, extending these efforts beyond the United States to countries such as the United Kingdom, Singapore, Malaysia, India, Japan, and South Korea. By leveraging their influence as foreign shareholders, they seek to promote greater gender diversity in boardrooms on a global scale (Csonka & Milhomem, 2023).

instance, the influence of individual shareholders in promoting board gender diversity is generally less effective than institutional shareholders. Although individual shareholders may actively submit proposals and pursue private negotiations with company management, their efforts frequently fail to achieve meaningful outcomes.

Research indicates that negotiations initiated by individual shareholders rarely lead to favourable agreements. Bauer et al. (2015) highlight the low rate of proposal withdrawals as evidence that individual shareholders often struggle to reach favourable resolutions.[9] This is primarily because individual shareholders lack the same level of access to information and expertise as institutional investors, which puts them at a disadvantage. Furthermore, another previous study argues that individual shareholders do not have the negotiation leverage or bargaining power that institutional investors possess, which further diminishes their ability to influence corporate decision-making (Rastad & Dobson, 2022). Without financial resources, insider knowledge, and voting power, proposals from individual shareholders often lack the credibility needed to be taken seriously by management.

Considering the preceding discussion, it is plausible to acknowledge the role of foreign institutional shareholders as a driver for board gender diversity. Therefore, I formulate the hypothesis as follows:

The presence of foreign institutional shareholders is expected to be positively associated with companies' decision to hire women on board.

In the upcoming section, I will discuss corporate governance, board gender diversity, and foreign ownership in an institutional setting with no mandatory regulations or quotas on female board members. As discussed in the earlier section, I will focus on Indonesia, a country known for its substantial number of women in the workforce yet low representation of women on boards. These characteristics create a unique context for analyzing the association in drivers that influence board member composition.

2.5. Institutional Setting

Before discussing board gender diversity in Indonesia, it is essential to understand the corporate governance in Indonesia. According to laws and regulations, Indonesia adopts a two-tier board structure: a supervisory board known as the Board of Commissioners (*Dewan Komisaris*), hereinafter referred to as "the BOC" and a management board known as the Board of Directors (*Dewan Direksi*), hereinafter referred to as "the BOD".[10] Public companies should have at least two board members within the BOC and the BOD. Among these members, one is designated as the president of the board. Additionally, the BOC shall have at least one or 30% independent members relative to the total member. Public companies are not required to appoint independent directors. Commercial banks have slightly different requirements regarding the composition of the boards. Commercial banks should have more than three board members on both boards, with at least 50% of the total BOC members serving as independent commissioners. In addition, Sharia banks and financial institutions adhering to Islamic law, known as Sharia Law, must have a Sharia Supervisory Board that monitors the company's product and activities under the Sharia Law.

All board members are appointed or dismissed by shareholders in the general meeting of shareholders (Rapat Umum Pemegang Saham). Before this meeting, with assistance from a Nominations and Remuneration Committee, the BOC provides recommendations, including but not limited to nominees for board members and compositions. Generally, board members shall serve the companies for a maximum of five years or until the annual general meeting of shareholders at the end of the fifth year. In addition, regulations prevent board members from holding dual roles within the same company. However, serving on boards across different companies, including those of controlling shareholders or parent companies, is permitted under laws and regulations except stated otherwise.[11]

To date, Indonesia lacks regulations addressing board gender diversity. As a result, public companies are still allowed not to have female board members on both boards. However, some public companies establish self-board diversity policies that align with their companies' values. For example, a large commercial bank has stipulated a female quota of two board members in the BOC, with one serving as an independent member and three board members in the BOD, with one serving as a president director.[12] In addition, the Indonesian Stock Exchange (IDX) has consistently shown commitment to promoting gender equality and board gender diversity.[13]

Nevertheless, despite the lack of specific regulations or mandates for female board members, Indonesia has witnessed increased representation of women on board over the past four years. As shown in Figure 2, the country had a representation of female directors below 5% before 2019 but rose to approximately 9% to 12% in the following years. Additionally, Graph B of Figure 3 highlights that in 2023, there is a slightly higher proportion of women in skilled roles, which include managerial positions, compared to men. Based on

[9] According to this study, individual shareholders' proposals have a withdrawal rate of around three to eight percents, while the proposals from institutional investors, coordinated activists, or labor unions have a withdrawal rate from 10 to 40 percents.

[10] As regulated in (i) Law No. 40 Year 2007 on Limited Liability Company, and/or (ii) Financial Services Authority Regulation No. 33 Year 2014 (POJK 33 2014) on the Board of Directors and the Board of Commissioners of Public Companies, and/or (iii) Financial Services Authority Regulation No. 55 Year 2016 (POJK 55 2016) on the Governance Implementation for Commercial Banks.

[11] See footnote 10. However, board members in commercial banks are prohibited from holding dual roles in other companies, except when serving as board members in non-financial subsidiaries controlled by the banks.

[12] As stated in the annual report of PT Bank CIMB Niaga Tbk.

[13] As an active member of the Sustainable Stock Exchanges (SSE), the IDX has consistently demonstrated its commitment, notably through its annual hosting of the "Ring the Bell for Gender Equality" conference since 2019.

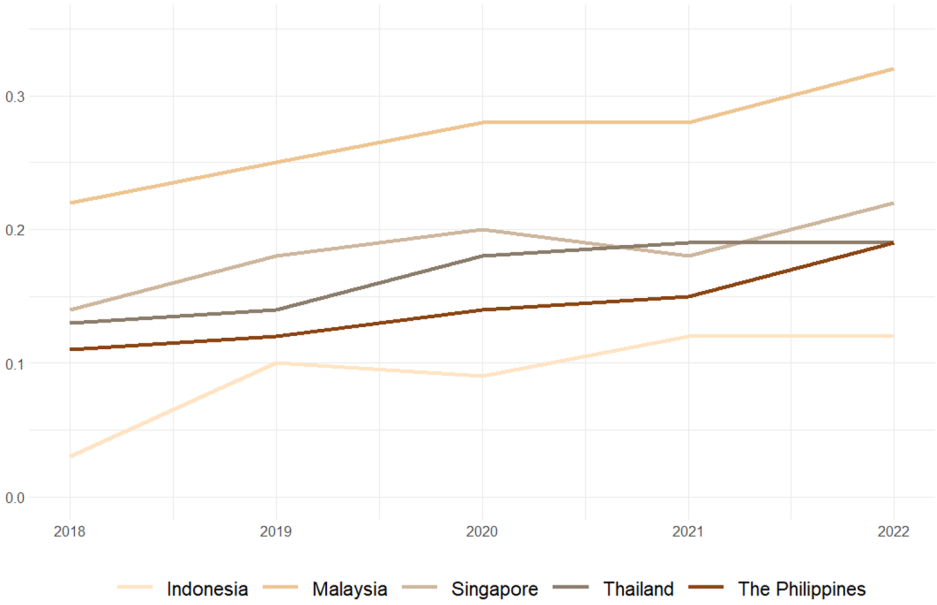


Figure 2: Proportion of Female Directors in Selected SEA Countries from 2018 to 2022

Annual averages of female directors based on data from public company boards. Source: Matanda et al. (2023).

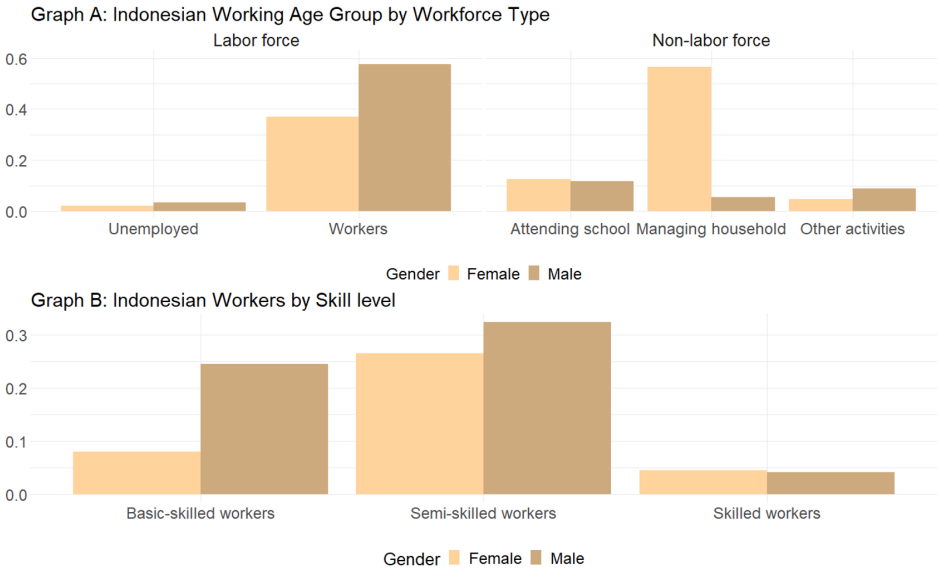


Figure 3: Gender Distribution in Indonesian Working Age Group by Workforce Type and Skill Level

Graph A shows the gender distribution among the working-age population engaged in the labor force and the non-labor force. Graph B shows the gender distribution by skill: skilled (professional, technical, managerial), semi-skilled (clerical, sales, service, agriculture, fishing), and basic-skilled (production, transport, labor). Classifications follow Indonesian Ministry of Labor guidelines. Source: BPS-Statistics Indonesia (2023) and The Indonesian Ministry of Labor (2023).

this evidence, it is plausible to infer that factors beyond legal regulations influence female leadership and representation in the boardrooms of Indonesian companies. However, the current representation of women on boards remains significantly low despite Indonesia's massive labor force of 148 million out of a 213 million working-age population.

Several factors may contribute to the underrepresentation of women in board positions in Indonesia. According to The Indonesian Ministry of Women Empowerment and Child Protection (2022), societal expectations often put women to manage the household rather than full-time employment, which is also reflected in Graph A of Figure 3. This graph reveals that only 36% of the labor force is female, and a majority of the non-labor force comprises women managing household duties. Moreover, there appears to be a bottleneck hindering Indonesian workers, including women, from attaining

executive roles. Graph B of Figure 3 illustrates that a large segment of the Indonesian workforce is concentrated in basic and semi-skilled jobs, which may limit the pool of female candidates for board positions.

Indonesia also falls behind neighboring countries in the context of board gender diversity. Figure 2 illustrates that the proportion of female executives in Indonesia is lower than in neighboring countries despite having the highest number of female workers. Further, some neighboring countries, such as Malaysia and Singapore, have implemented female board regulations for their public companies. Meanwhile, Thailand and the Philippines, even without board gender diversity mandates, still have a higher proportion of female board members than Indonesia.

In addition, Indonesia is not only falling behind in regulatory measures and female board representation but also in the academic literature on board gender diversity. The limited research predominantly examines the relationship between board gender diversity and corporate performance, yielding mixed results. A prior study highlights a positive link between women in executive roles and the performance of Indonesian financial firms and institutions (Tjondro et al., 2020). Similarly, another previous study reports a positive correlation between female board representation and gas emission disclosures among Indonesian public companies (Daromes & Djie, 2019). In contrast, a study identifies a negative correlation between female board membership and firm performance within 169 IDX-listed Indonesian firms in 2007 (Darmadi, 2011). Moreover, a previous study finds no association between female board presence and ESG performance in Indonesian public firms (Brown et al., 2023). Additionally, there seems to be a shortage of studies investigating drivers of board gender diversity in Indonesia.

While Indonesia faces challenges in board gender diversity, its economy stands out as one of the world's strongest economies. The country ranks as the largest economy in Southeast Asia (SEA) and the 16th largest economy in the world, with an estimated GDP of 19.5 trillion rupiah, which translates to 1.3 billion US dollars (IMF, 2023). Further, the country attracts substantial foreign direct investment due to its vast consumer market driven by a population of 275 million. Indonesia notably has received an inflow of more than one billion USD for foreign investments from Singapore, China, Japan, Hong Kong, The Netherlands, Malaysia, South Korea, and the US (NSWI, 2023).

Likewise, Indonesia has attracted a significant deal of foreign investment within its capital market. Notably, from 2010 to 2019, the country recorded an average annual net foreign investment of 89.83 trillion rupiah (Deloitte Indonesia, 2021). Further, the Financial Services Authority highlighted a continued interest in the Indonesian stock market, with foreign investors contributing a net inflow of 2.72 trillion rupiah as of July 2023 (Financial Services Authority, 2023). This evidence highlights the consistent presence of foreign shareholders in Indonesia's financial markets, suggesting their potential influence on corporate decisions, including board gender composition. Therefore, it is important to explore the relationship between foreign shareholders and board gender diversity among companies in Indonesia.

3. Data and Methodology

3.1. Sampling Method and Distribution

To empirically answer the research question, I will focus on female representation among Indonesian companies' boards between 2019 and 2022. As discussed in the previous section, the percentage of female board members was substantially low before 2019. Further, advocacy for board gender diversity in Indonesia was relatively scarce before 2019.

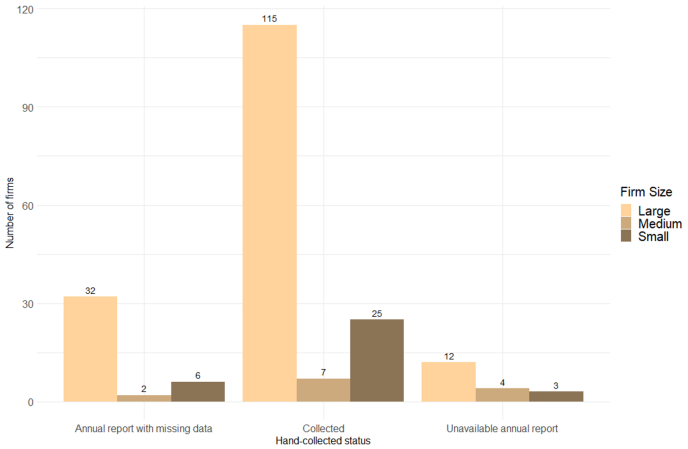
I define the study population as public companies listed in the IDX. I retrieve the data for Indonesian public companies from the Refinitiv Worldscope database, which contains detailed financial information and company profiles of public companies from around the world. Subsequently, I identify 902 companies listed on the IDX using their ISIN codes and exclude 257 companies as they are listed after 2019, thus not fulfilling the sample period coverage that I have set previously. Therefore, the sample pool comprises 645 Indonesian public companies.

To obtain the sample, I employ a stratified random sampling method. This method requires grouping the data with similar characteristics, known as the strata. These strata are then randomly sampled using a different probability sampling method. I categorize the strata according to the 11 industry sectors from the Global Industry Classification Standard (GICS), then set a sample size of 20 companies for each stratum. The sample size reflects consideration of the time constraints while ensuring the representativeness of the population within the sample. However, one industry sector has only six companies within the sample period. Thus, I exclude this sector from the sampling pool and included all six companies as the sample. Consequently, this process generates a final sample of 206 Indonesian public companies.

As shown in Figure 4, the outcome of the stratified random sampling process is primarily concentrated in large firms, directly reflecting the majority of Indonesian public companies categorized as large firms. Meanwhile, medium and small firms comprise 16% and 6% of the sample. However, during the manual data collection process, as outlined in this figure and Table 1, the data for some of these firms are uncollectible. A notable issue is the absence of publicly disclosed shareholder classifications by domestic or foreign status, which is necessary for measuring foreign institutional investors. This issue suggests that some public companies are not complying with the regulations. Another issue is the lack of accessible annual reports on the company's and IDX's websites. This challenge is especially notable among companies in the Communication Services, Health Care, Industrials, Information Technology, and Real Estate sectors, as shown in Table 1. It also affects large entities, with data for many large firms being uncollectible. As a result, the number of sample firms is reduced from 206 to 147.

Figure 4: Sample Distribution by Firm Size and Hand-Collected Data Status

Firm Size	Population Firms	Initial Sample Firms
Large	639	159
Medium	130	13
Small	121	34
Unavailable firm-size data	12	0
Total	**902**	**206**



The table shows the distribution of initial sample firms (2019–2022) by size, classified as small, medium, or large based on total assets in the 2022 annual report. Source: (Refinitiv, 2022); see Table 2 for definitions of firm size. The bar plot shows the same firms by size and hand-collected data status: Collected (annual reports with usable data), Annual report with missing data (no shareholder classification: foreign/domestic), and Unavailable annual report (no online report for the period).

Table 1: Sample Distribution by Industries and Hand-Collected Data Status

GICS Sector Name	Population	Initial Sample Firms	Hand-collected data status		
			Collected (Final Sample Firms)	Annual report with missing data	Unavailable annual report
(1)	(2)	(3)	(4)	(5)	(6)
Energy	56	20	16	4	0
Materials	99	20	17	2	1
Industrials	145	20	11	5	4
Consumer Discretionary	117	20	16	3	1
Consumer Staples	116	20	16	3	1
Health Care	32	20	13	6	1
Financials	109	20	16	3	1
Information Technology	38	20	10	3	7
Communication Services	48	20	13	7	0
Utilities	8	6	5	1	0
Real Estate	91	20	14	3	3
Unavailable GICS code data	43	0	0	0	0
Total	**902**	**206**	**147**	**40**	**19**

This table classifies Indonesian public firms by GICS industry sector, plus one group for firms without a GICS code. Column (2) lists the population of public firms listed in the Indonesian Stock Exchange (IDX) as of December 2023. Column (3) shows the initial stratified random sample by sector. Columns (4)-(6) report firms with data successfully gathered from annual reports, firms whose reports lacked shareholder classification (foreign/domestic), and firms with no available annual reports, reducing the final sample to 147 firms.

Table 2: Variable Definitions

	Variable name	Description	Source
Dependent variables	% FC	Percentage of female members in the Board of Commissioners.	Annual report
	Dummy FC	A dummy variable that is equal to 1 if a firm has at least one female commissioner and 0 otherwise.	
	% FD	Percentage of female members in the Board of Directors.	
	Dummy FD	A dummy variable that is equal to 1 if a firm has at least one female director and 0 otherwise.	
Independent variable	% FII	Percentage of foreign institutional investors.	
Control variables	BOC Size	The size of the Board of Commissioners is defined as the total number of board members. For public companies, the minimum board size is two members, with one member serving as President and at least 30% serving as independent commissioners. For commercial banks, the minimum is three members, with at least 50% serving as independent commissioners.	
	BOD Size	The size of the Board of Directors is defined as the total number of board members. The minimum required size is two members for public companies and three for commercial banks, where one member acts as President. Independent directors are not required in either public companies or commercial banks.	
	%LII	Percentage of local institutional investors.	
	Firm Size (ln)	Defined as the natural logarithm (ln) of the book value of assets. According to Financial Services Authority Regulation No. 53 Year 2017 (POJK 53 2017) on Initial Public Offering for Small and Medium Enterprises, small firms have total assets of not more than IDR 50 billion, medium firms have total assets between IDR 50-250 billion, and large firms have total assets of more than IDR 250 billion.	Refinitiv Eikon

3.2. Research design and descriptive statistics

I define the dependent variable as the presence of women on boards and the independent variable as foreign institutional investors. I hand-collect the data from the annual reports due to the unavailability of accessible information on secondary databases. Furthermore, I introduce additional control variables to mitigate potential endogeneity arising from omitted unobservable factors, which could otherwise result in spurious correlations. The control variables are board size, firm size, and local institutional shareholders. I will retrieve the data for these variables from the annual reports and Refinitiv Eikon. Previous studies on board gender diversity have frequently utilized board size and firm size as control variables, suggesting that the size of the board and the firm may affect the likelihood of including women on the board (Adams & Ferreira, 2009; Fauver et al., 2022; Rastad & Dobson, 2022). In addition, I will use a two-way fixed effects model in the analysis to control for time-variant and time-invariant unobserved variables that might occur in the data.

To examine the hypothesis, exploring the association between the presence of women on boards and foreign institutional shareholders while controlling for endogeneity, I define the regression model as follows:

$$\begin{aligned}\text{Women on boards}_{it} = \\ \beta_0 + \beta_1 \text{Foreign institutional investors}_{it} \\ + \beta_2 \text{Board size}_{it} + \beta_3 \text{Firm size}_{it} \\ + \beta_3 \text{Local institutional investors}_{it} + \alpha_i + \gamma_t + \epsilon_{it}\end{aligned}$$

This study aims to examine how the influence of foreign shareholders on board gender diversity differs in companies with a two-tier board structure, where supervisory (BOC) and management (BOD) functions are separated, and

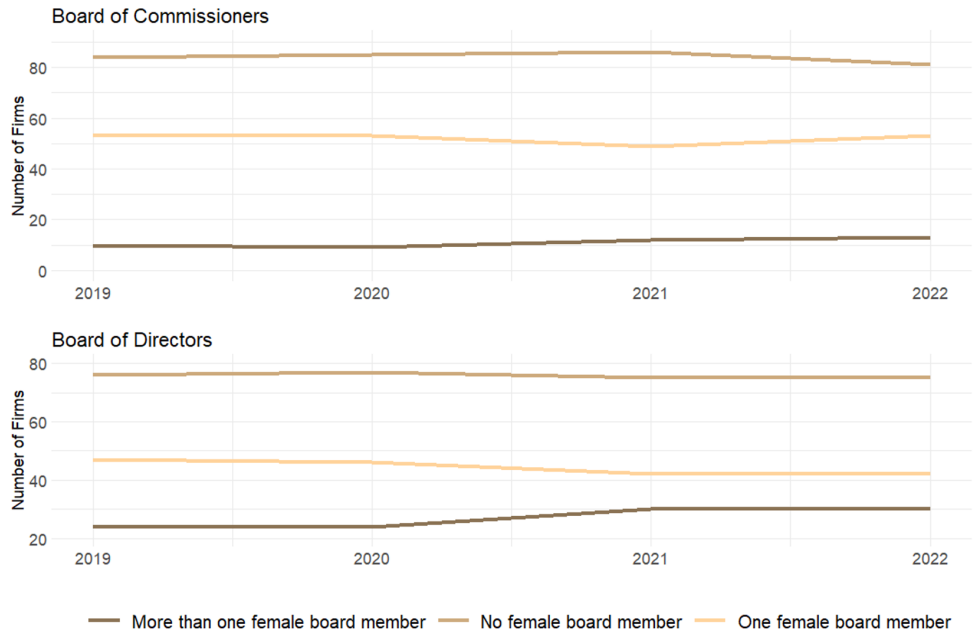


Figure 5: Female Compositions on Board from 2019 to 2022

This figure shows the number of firms with no female board members, those with one female board member, and those with more than one female board member for each year, respectively. The top graph displays these compositions within the Board of Commissioners across the observed period, while the bottom graph depicts the same for the Board of Directors.

whether foreign investors exert the same level of influence on both boards within this governance model. While much of the existing literature focuses on foreign shareholders' impact on gender diversity in single-board systems, the unique dynamics of a two-tier structure require further exploration. To address this, I measure the presence of women on both the BOC and BOD by calculating the percentage of female board members on each board. Additionally, I introduce two dummy variables for the dependent variable—one for the BOC and another for the BOD—where each variable is assigned a value of 1 for firms with at least one female member on the respective board and 0 otherwise.

The primary independent variable, foreign institutional shareholders, is measured by the percentage of foreign institutional investors in the firm. To control for other factors that might influence board composition, I include board size, defined as the total number of members on each board; firm size, defined as the logarithm of the book value of assets; and local institutional shareholders, defined as the percentage of local institutional investors in the firm. Given the different units of value in the dependent variables, I employ an OLS regression model for the percentage of female board members and a probability linear model for the dummy variables. These models allow me to explore whether foreign shareholders exert a differential impact on gender diversity in the supervisory and management boards, addressing the gaps identified in the previous literature.

Figure 5 presents the board gender composition of the BOC and the BOD among the sample firms from 2019 to 2022. This figure shows that firms without female representation on both boards outnumber those that include female members. Further, year-over-year trends reveal a relatively stable presence of all-male boards, with a slight reduction in such boards within the BOC towards the end of the period. Meanwhile, firms with at least one female board member show slight fluctuations, particularly in the last two years. A notable observation is the decrease in firms with exactly one female director in 2021. However, this trend can be linked to the increase in firms with more than one female director, suggesting a growing presence of women in the BOD.

Similarly, firms with one female commissioner show a slight decline from 2020 to 2021 but rebound in 2022, while firms with more than one female member in the BOC show a subtle yet consistent growing trend. Further, although the data suggest that the board gender composition among the sampled firms has not notably changed, a modest gap persists between firms with no female board members and those with at least one. Approximately 41% to 49% of the firms consistently included at least one female board member throughout the sample period, reflecting moderate awareness and implementation of gender diversity within Indonesian companies.

Figure 6 demonstrates the proportion of sample firms with at least one female board member, segmented by firm size. It shows that, between 2019 and 2022, 51% to 55% of large firms have at least one female director, while approximately 43% to 47% have at least one female member in the BOC. This trend aligns with previous studies that suggest large firms tend to have more female board members (Hillman et al., 2007; Rastad & Dobson, 2022; Shams, 2019). Large firms are the most visible to the public and have more

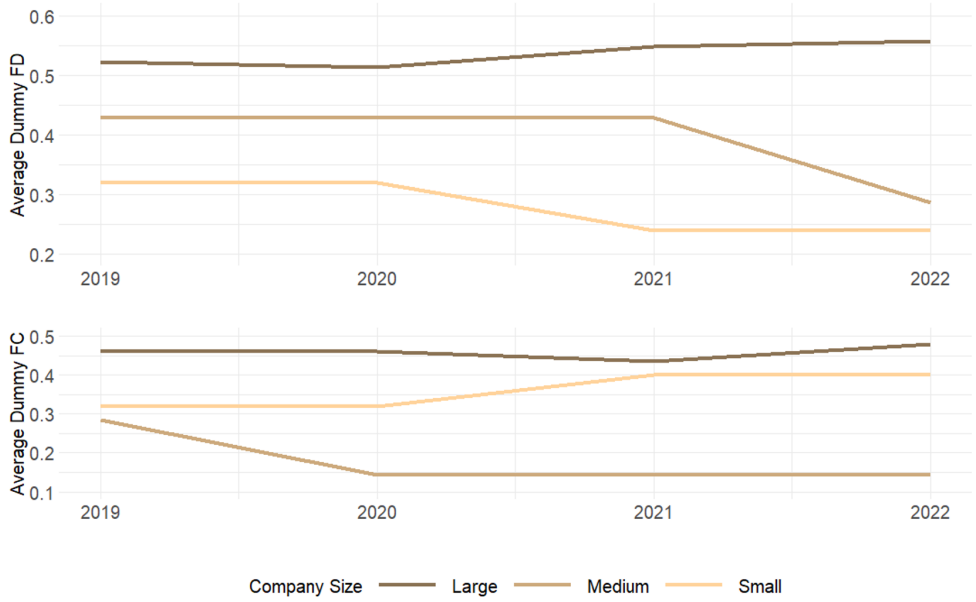


Figure 6: Average Firms with Dummy Female Board Members by Firm Size

This figure presents the annual percentages of firms with at least one female member (dummy = 1) and firms without (dummy = 0), categorized by firm size and board type.

pressure to increase female representation on boards (Hillman et al., 2007). Moreover, they are usually associated with a larger board size, which indicates a potential capacity to include more diverse members, although this does not automatically guarantee such diversity. Despite this, the trend from 2019 to 2022 shows slight fluctuations, corroborating a lack of significant change in board gender diversity, as observed in Figure 5.

Interestingly, the trend for medium and small firms varies between the BOC and the BOD. Small firms show a higher percentage of female commissioners than medium firms, while medium firms tend to have female directors than small firms. This distinction might suggest that small firms prioritize enhancing their governance practices by utilizing women's strong monitoring abilities, particularly on a supervisory board. According to a study by Adams and Ferreira (2009), women are often appointed to boards for their monitoring capabilities, and firms with weaker governance practices might benefit more from having women on the board. Therefore, the strategic emphasis on governance improvement may drive the appointment of female commissioners in small firms. Meanwhile, medium firms might focus on operational or strategic roles, as reflected by the higher presence of female directors.

Interestingly, the data point out that there are more women on the BOD than on the BOC throughout the sample period, as illustrated in Figure 7. This insight is further emphasized in Table 3, where the average proportion of female directors is higher than that of female commissioners. However, while the overall female representation is higher on the BOD, women more frequently occupy the highest position on the board, i.e., president, in the BOC from 2019 to 2022. Despite this representation, only 13% of BOC presidents and 9% of BOD presidents are women within the 588 firm-year observations. This proportion indicates the scarcity of female members in this leading role, which could be associated with the limited number of women within the candidate pool. In addition, Figure 7 reveals another insight regarding women's role in other board positions. It shows that over 30% of women in the BOC serve as independent members. Similarly, a prior study by Abdullah (2014) finds that approximately 25% of women are independent members.[14] As discussed earlier, women have stricter monitoring capabilities, which may influence their selection as independent members within boards (Adams & Ferreira, 2009).

Table 3 details the descriptive statistics of all the variables, accounting for 588 firm-year observations. The data suggest positive skewness, as indicated by the means exceeding the medians, in the dependent and independent variables. This skewness occurs due to a large number of the sample firms having little to no female representation on each board or foreign institutional investors, which is depicted further in the scatter plots in Figure 8. Consequently, the proportion of female board members and foreign institutional investors often falls below the first percentile for these variables, resulting in positive skewness. Considering these ratio variables inherently exhibit limited variation, and given the

[14] As discussed in the previous chapter, the appointment of independent directors is not mandatory, resulting in a significantly low presence of such role among the sample firms. Consequently, including this role in the analysis is unlikely to yield meaningful insights and thus, I exclude it from the analysis.

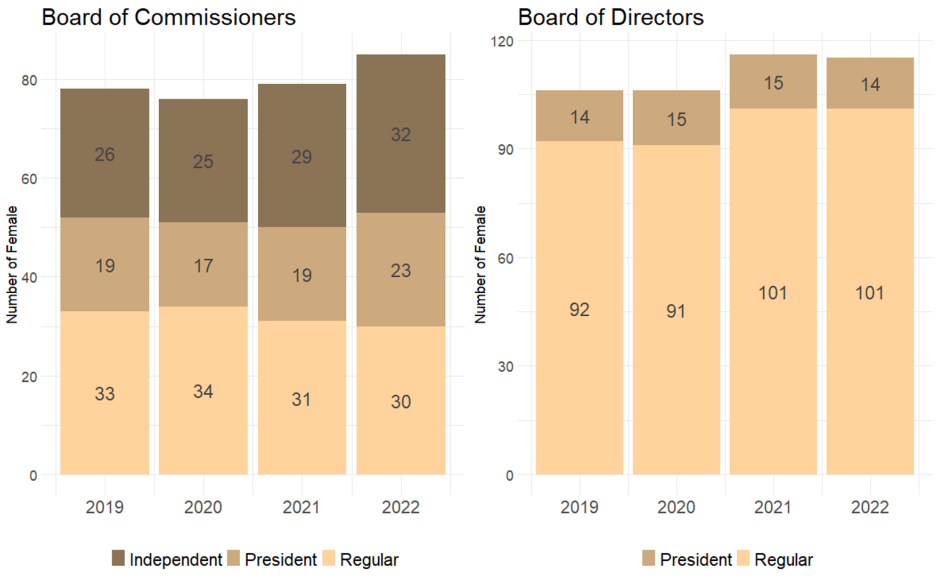


Figure 7: Female Board Members by Position between 2019 and 2022

This figure presents the annual count of female board members by position across 147 sample firms. Each stacked bar shows the number of female members per position, summing to the total female board count each year. Unlike the Board of Commissioners, the Board of Directors is not required to have independent members but must appoint a president. Note that the y-axis scales differ to accommodate data range variations.

Table 3: Descriptive statistics

Variable	N	Mean	Std. dev.	Min.	Q1	Median	Q3	Max.
%FC	588	0.136	0.183	0	0	0	0.250	1
Dummy FC	588	0.429	0.495	0	0	0	1	1
%FD	588	0.168	0.212	0	0	0	0.333	1
Dummy FD	588	0.485	0.500	0	0	0	1	1
%FII	588	0.250	0.283	0	0.023	0.141	0.349	0.997
BOC Size	588	4.04	1.99	2	3	3	5	16
BOD Size	588	4.49	1.92	2	3	4	5	11
Firm Size (IDR Million)	588	13,210,801	41,204,153	62	275,599	1,940,098	7,611,470	413,297,000
Firm Size (ln)	588	27.402	3.512	17.935	26.342	28.294	29.661	33.655
%LII	588	0.527	0.307	0.000	0.270	0.588	0.790	0.999

This table provides the summary statistics of the variables for 147 firms, encompassing 588 observations from 2019 to 2022. Within the firm-year observations, the BOC is comprised of 2,375 members, while the BOD includes 2,638 members. For details on variables' definitions, see Table 2.

reasons for the observed skewness, I do not address the skewness to preserve the natural variability of the data. However, due to significant variability within the range of value of firm size, I address the skewness in this variable using natural logarithm.

The data reveals that, among the 2,375 members serving on the BOC, females constitute an average of 13.6%, with a standard deviation of 18.3%. In comparison, within the BOD, females represent an average of 16.8% of the 2,638 members, with a standard deviation of 21.2%. These statistics suggest a modest level of female representation on the boards of the sampled firms. Additionally, the data show an average presence of foreign institutional investors at 25%, with a standard deviation of 28.3%, indicating that foreign investors have a lower presence compared to domestic investors, who account for an average of around 52% among the sample firms. However, it's important to approach these results with caution due to the unique characteristics of each firm. This implies that a generalized average value might not fully capture meaningful trends in board gender diversity and

Table 4: Correlation Coefficients

Panel A: Correlation coefficients (Female Commissioners)

	% FC	Dummy FC	% FII	BOC Size	Firm Size (ln)	% LII
% FC	1		0.026	0.044	0.127***	-0.011**
Dummy FC		1	0.065	0.188***	0.166***	-0.09**
% FII	0.001	0.056	1	0.265***	0.104**	-0.60***
BOC Size	-0.025	0.198***	0.203*	1	0.360***	-0.03
Firm Size (ln)	0.106**	0.129***	-0.057	0.202***	1	0.11***
%LII	-0.124***	-0.09**	-0.650***	-0.039	0.146***	1

Panel B: Correlation coefficients (Female Directors)

	% FD	Dummy FD	% FII	BOD Size	Firm Size (ln)	% LII
% FD	1		0.031	0.073*	0.163***	0.02
Dummy FD		1	0.083**	0.209***	0.189***	0.00
% FII	0.035	0.123***	1	0.343***	0.104*	-0.60***
BOD Size	-0.005	0.228***	0.314***	1	0.464***	-0.04
Firm Size (ln)	0.167***	0.200***	-0.057	0.341***	1	0.11***
%LII	0.051	-0.021	-0.650***	-0.094**	0.146***	1

This table presents Pearson (below diagonal) and Spearman (above diagonal) correlation coefficients. Panel A uses female commissioners (FC) as the dependent variable, while Panel B uses female directors (FD). Each variable is measured by two proxies: percentage and a dummy (1 if firm has at least one female board member, 0 otherwise). The main independent variable is the percentage of foreign institutional investors (% FII). Control variables include board size, natural logarithm of firm size, and percentage of local institutional investors (% LII). For simplicity, the p-values are not reported but indicated by the significance levels as (***) $p < 0.01$, (**) $p < 0.05$, (*) $p < 0.1$.

foreign institutional investor involvement. Therefore, I will discuss the association between these variables in greater detail following a series of more robust analyses.

Table 4 presents scatter plots and correlation coefficients for all variables, calculated using both Pearson and Spearman methods. These correlations help to explore the general correlation between foreign and local institutional investors and board gender diversity, allowing us to understand the pattern of the data to elicit the association between the two variables.

Panel A reveals a weak and non-significant correlation between the percentage of female commissioners and foreign institutional investors, as indicated by both Pearson and Spearman's methods. It is probable that foreign investors might not have a strong influence on increasing the percentage of women on the BOC. Additionally, Panel A of Figure 8 shows no clear linear association between these variables, with scattered data points and no apparent trend. This supports the earlier finding of a potential lack of influence from foreign institutional investors on board gender diversity in BOC. On the other hand, as shown in Table 4, local institutional investors exhibit a significant but weak negative correlation with both the percentage of female commissioners and the dummy variable for female commissioners. This negative correlation supports the earlier notion that domestic local investors might be less supportive of board gender diversity, possibly due to a lack of awareness and limited perspective on this matter.

The coefficients between firm size and female commissioners show a positive and significant correlation in both Pearson and Spearman methods, indicating that larger firms are more likely to have female commissioners. However, while board size is positively correlated with the likelihood of having at least one female commissioner, this does not extend to the overall percentage of female representation in the BOC. These trends are visually depicted in Panel A of Figure 8, where the fitted regression lines show different relationships between board size and the dummy versus percentage variables.

Panel B of Table 4 reveals positive and significant correlations between foreign institutional investors and the dummy variable for female directors (12.3% for Pearson and 8.3% for Spearman). This indicates that foreign institutional investors are associated with a higher likelihood of having at least one female director on the board of directors (BOD). The upward trend of the fitted regression line in Panel B of Figure 8 further supports this association. However, this positive correlation does not extend to the overall percentage of female directors, suggesting that foreign investors may focus on ensuring minimal female representation rather than

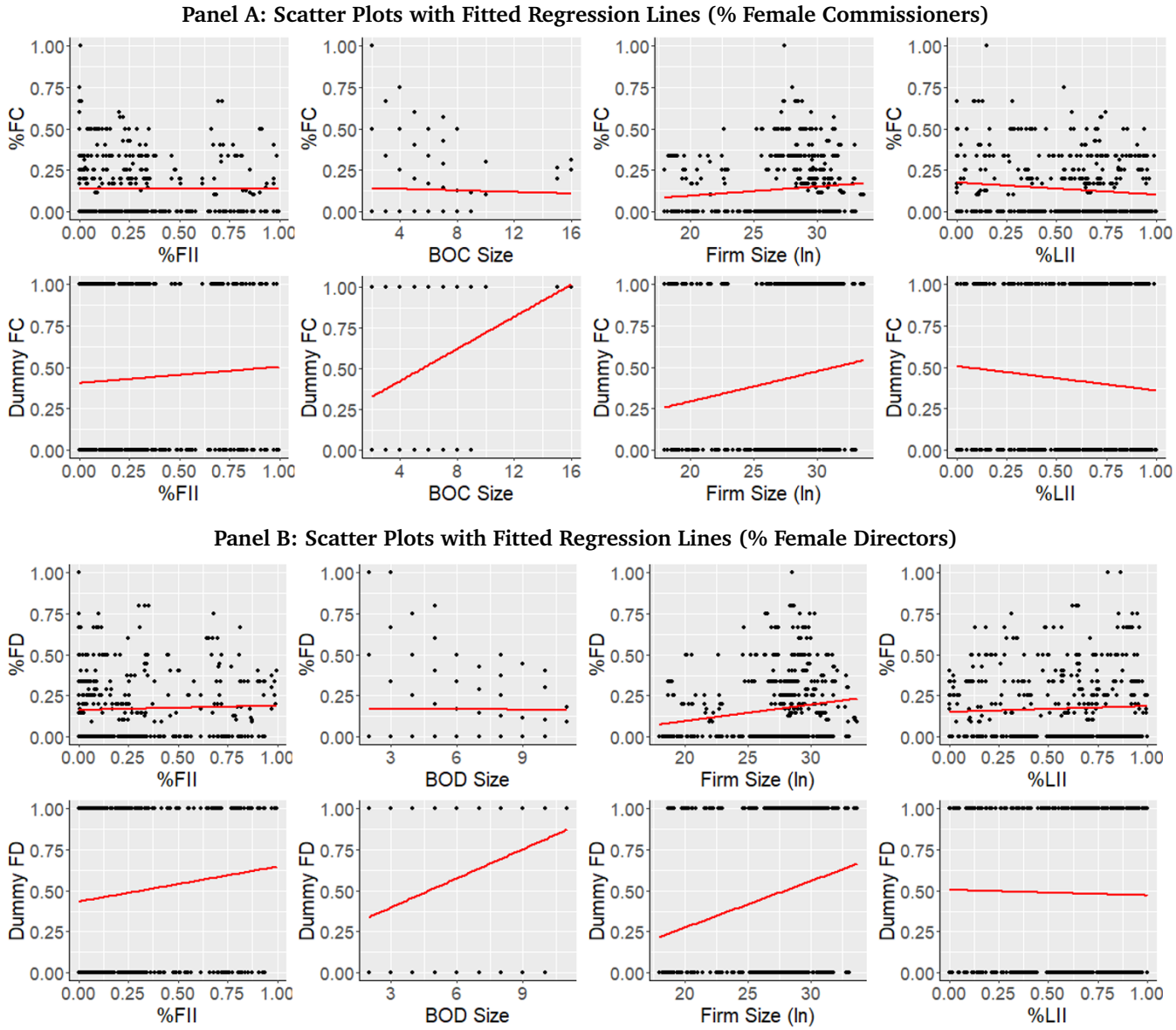


Figure 8: Scatter Plots

Panel A shows scatter plots between key variables using female commissioners (FC) as the dependent variable, while Panel B using female directors (FD). Each variable is measured by two proxies: percentage and a dummy (1 if the firm has at least one female board member, 0 otherwise). The main independent variable is the percentage of foreign institutional investors (% FII). Control variables include board size, the natural logarithm of firm size, and the percentage of local institutional investors (% LII).

a proportional increase in gender diversity. Interestingly, unlike in Panel A, no significant correlation is found between the percentage of local institutional investors and the likelihood of having a female director or the overall rate of female directors. This finding suggests that, unlike foreign institutional investors, local institutional investors are less likely to influence the presence of female directors on the board.

In general, these findings highlight nuanced differences in how foreign and local institutional investors influence board gender diversity. Foreign investors show some association with the presence of female directors, while local investors may exhibit more resistance. However, these results are based on simplified relationships between the dependent and independent variables, offering limited information on potential causal inferences. To gain deeper insights and address these limitations, it is crucial to apply a more robust analysis, such as regression modeling, to understand better the underlying dynamics and control for potential confounding factors.

The comprehensive trend analysis, as illustrated in Table 5, shows annual variations among the observations for the key dependent and independent variables. Panel A indicates noticeable cross-sectional variability, with standard deviations ranging from 17% to 29%. This suggests that firm-

Table 5: Comprehensive Trend Analysis of Female Board Composition and Foreign Institutional Shareholders from 2019 to 2022

Panel A: Annual Descriptive Statistics Across All Sample Firms for Each Year

	%FC		%FD		%FII	
Year	**Mean**	**Std. dev.**	**Mean**	**Std. dev.**	**Mean**	**Std. dev.**
2019	0.136	0.179	0.160	0.202	0.259	0.292
2020	0.133	0.177	0.163	0.208	0.246	0.280
2021	0.132	0.184	0.174	0.221	0.242	0.278
2022	0.144	0.191	0.176	0.219	0.252	0.283

Panel B: Year-Over-Year Trends in %FC, %FD, and %FII for Firms from 2019 to 2022

%FC \ %FH	Decrease	Increase	No change	**Total**
Decrease	8	12	9	**29**
Increase	21	20	4	**45**
No change	167	107	93	**367**
Total	**196**	**139**	**106**	**441**

%FD \ %FH	Decrease	Increase	No change	**Total**
Decrease	15	14	6	**35**
Increase	25	12	14	**51**
No change	156	113	86	**355**
Total	**196**	**139**	**106**	**441**

This table examines the dynamics of female board composition and foreign institutional investors across 147 firms from 2019 to 2022. Panel A presents cross-sectional annual means and standard deviations, showing year-by-year distributions. Panel B provides a panel data analysis comparing changes in female commissioners (%FC) or directors (%FD) with foreign institutional investor percentage (%FII).

specific characteristics, such as firm performance as well as governance quality, influence the representation of female board members. Further, this panel also provides insight into the annual trend of the key-dependent and independent variables. While the mean percentage of female commissioners and directors shows a modest upward trend, foreign institutional investors demonstrate a marginal decline. Such temporal fluctuations may be attributable to factors that are consistent across firms but vary over time, such as changes in regulations or economic conditions. Nonetheless, the presence of both cross-sectional and time variations underscores the potential for unobserved heterogeneity that could impact the estimated relationships in the regression model, which should be addressed to mitigate endogeneity. This consideration is consistent with the earlier assumption of the regression model and highlights the importance of including firm-year fixed effects in the regression model.

Panel B of Table 5 illustrates the relationship between the aggregate changes in the key-dependent and independent variables across the firm-year observations. The association between them appears to be obscure, indicating that variations in the percentage of foreign institutional investors are not consistently associated with shifts in the percentage of female commissioners and directors. This aligns with the weak and non-linear correlations previously identified in Table 7 and Figure 8. Moreover, the data in this table reinforce the patterns observed in Figure 5, which suggests a persistent trend of all-male board compositions. This trend is reflected in the minimal number of firms showing increases or decreases in the percentage of female commissioners and female directors.

Interestingly, Panel B also indicates that the sample firms more frequently experience a decrease rather than an increase in foreign institutional investor composition. As shown in Panel A, there are slight decreases in the percentage of foreign investors in 2020 and 2021, which potentially correlates with the COVID-19 pandemic that hinders global investors from investing in foreign countries. This implica-

tion is also highlighted by Deloitte Indonesia (2021), which reports a notable withdrawal of foreign investments in Indonesia's capital market, amounting to 135 trillion rupiah in 2020.

Overall, Panel B reveals no discernible association between female board members and foreign institutional investors. However, I will further investigate this relationship through regression analyses, offering a more comprehensive exploration of the association.

4. Results and Discussions

4.1. Regression Results and Discussions

Table 6 details the regression results, showing the association between institutional investors and female representation on the Board of Commissioners (BOC) and Board of Directors (BOD), both with and without two-way fixed effects. Robust standard errors adjusted for heteroscedasticity, enhancing the reliability of the estimates.

The analysis indicates a significantly positive association between foreign institutional ownership and female representation on the BOD, as shown in columns (5) and (7). Specifically, a higher percentage of foreign investors correlates with a greater likelihood of having at least one female director and a higher overall proportion of female directors. This supports the hypothesis that foreign institutional investors value gender diversity, particularly in management roles. However, this positive relationship does not extend to the BOC, where an increased proportion of foreign investors is associated with a decrease in female commissioner representation (column 1). This suggests that foreign investors prioritize gender diversity in executive roles over supervisory roles, potentially viewing female directors as more critical for strategic decision-making and governance, as previously highlighted by Adams and Ferreira (2009).

However, the significance of these associations weakens after incorporating firm-year fixed effects, indicating that unobserved time-invariant factors may significantly influence board composition beyond the impact of foreign ownership alone. This is consistent with the relatively small proportion of foreign institutional investors (25%) compared to local counterparts (52.7%) in the sample firms, as indicated in Table 6. This indicates that foreign institutional investors may have a limited impact on board gender diversity. In contrast, local institutional investors could exert a more significant influence due to their familiarity with local business practices. It is reasonable to assume that local institutional investors might prioritize factors such as enhancing financial performance over promoting board gender diversity, especially given the relatively low awareness of this issue. Additionally, although one coefficient remains significant, the association turns negative. This finding indicates that a higher proportion of foreign institutional investors is associated with a reduced likelihood of firms appointing female directors, contrasting the original hypothesis. Further, this finding contrasts with the results of Fauver et al. (2022), who discovered that foreign institutional investors continue to promote board gender diversity, even when accounting for two-way fixed effects.[15]

Turning to local institutional investors, the results show mixed associations with female representation on the BOC and BOD, as shown in Columns (1) and (3). A one percentage point increase in local institutional investors is associated with a decrease of −0.139 percentage points in the percentage of female commissioners and a decrease of −0.234 percentage points in the likelihood of having at least one female commissioner. This suggests local institutional investors might prioritize other agendas over board gender diversity in commissioner roles. Conversely, a positive association is observed between local institutional investors and the proportion of female directors, indicating their role in promoting female representation at the director level. This complexity highlights that promoting gender diversity may require efforts from both foreign and local investors. However, after including two-way fixed effects, these associations largely diminish, indicating that the observed positive relationships may have been influenced by unobserved factors rather than direct actions by local investors.

Column (5) shows a weak but significant negative relationship between BOD size and the proportion of female directors, indicating that an increase in BOD size is associated with a decrease of 0.013 percentage points. In contrast, columns (3) and (7) reveal a positive relationship between board size and the likelihood of having at least one female member on the BOC and BOD. This finding is consistent with the results reported, which aligns with previous results by Abdullah (2014), Fauver et al. (2022), and Shams (2019). After controlling for firm-year fixed effects, the positive association for BOD size remains significant, indicating that larger BODs are still more likely to have female representation, even though the overall proportion of female directors may not necessarily increase. In other words, perhaps the inclusion of females in the BOD within the sample firms is merely a gesture of aligning public interests on board gender diversity, consistent with Figure 5, where most sample firms have none to one female directors.

Additionally, the coefficient results reveal that firm size is positively associated with the proportion of female members and the likelihood of having at least one female member on the BOC and BOD. Although these associations are weak, they suggest that larger firms tend to have a higher representation of female members. However, none of these associations persist after incorporating two-way fixed effects, which may indicate that other unobserved factors could influence the observed relationships. This finding underscores the need for further investigation into the factors that drive female representation on boards.

The regression analysis reveals nuanced insights into the relationship between foreign institutional investors and female board representation on the Board of Commissioners (BOC) and the Board of Directors (BOD). While the initial

[15] This discrepancy might arise from their focus on lagged foreign institutional ownership, which may capture different dynamics.

Table 6: Regression Results

	Dependent Variables: Female Commissioners				Dependent Variables: Female Directors			
	%FC		Dummy FC		%FD		Dummy FD	
Explanatory Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
%FII	-0.089**	-0.065	-0.124	-0.014	0.117***	-0.046	0.198**	-0.270*
	(0.041)	(0.29)	(0.097)	(0.289)	(0.039)	(0.039)	(0.098)	(0.143)
BOC Size	-0.003	0.002	0.045***	0.101				
	(0.003)	(0.013)	(0.009)	(0.033)				
BOD Size					-0.013***	0.001	0.038***	0.093***
					(0.004)	(0.008)	(0.012)	(0.024)
Firm Size (ln)	0.007***	0.005	0.015***	-0.016	0.012***	-0.006	0.021***	-0.019
	(0.002)	(0.007)	(0.006)	(0.012)	(0.003)	(0.005)	(0.006)	(0.014)
%LII	-0.139***	0.030	-0.234***	0.034	0.078**	-0.100**	0.072	-0.034**
	(0.038)	(0.061)	(0.091)	(0.220)	(0.038)	(0.040)	(0.090)	(0.0142)
N Observations	588	588	588	588	588	588	588	588
R-squared	0.04	0.8693	0.060	0.8418	0.045	0.91	0.075	0.88
Adjusted R2	0.04	0.8232	0.054	0.7861	0.0387	0.89	0.070	0.845
Two-way Fixed Effects	No	Yes	No	Yes	No	Yes	No	Yes

This table presents linear regression results with two-way fixed effects for a sample of 147 firms from 2019 to 2022 (588 observations). Columns are grouped by dependent variable: (1)-(2) percentage of female commissioners (%FC), (3)-(4) dummy for female commissioners, (5)-(6) percentage of female directors (%FD), and (7)-(8) dummy for female directors. Dummy variables equal 1 if a firm has at least one female board member, 0 otherwise. All variables use an OLS regression model, except for models with dummy dependent variables, which use a linear probability model. The main independent variable is the percentage of foreign institutional investors (%FII), with controls for percentage of local institutional investors (%LII), board size, and firm size. Coefficients are shown with robust standard errors in brackets. Significance levels: (***) p<0.01, (**) p<0.05, (*) p<0.10.

hypothesis predicted a positive association between foreign ownership and female board representation, the results reveal a more complex relationship, with positive effects on the BOD but negative effects on the BOC. However, these associations largely dissipate after controlling for two-way fixed effects, suggesting that foreign and local institutional investors may not be the primary drivers of gender board diversity among the sample firms and that other factors may exert more influence.

4.2. Endogeneity Test: Regression with Lead Model Specification

The earlier analyses controlled for endogeneity by utilizing control variables and two-way fixed effects. However, endogeneity may also occur if the assumed direction between dependent and independent variables is reversed, also known as reverse causality. Failing to account for reverse causality can lead to biased estimates in the regression models (Adams, 2016).

One method to control for reverse causality is to use a lead model specification. This approach assumes that the current value of the independent variable affects future values of the dependent variable rather than the reverse. By including future values of the dependent variable (leads) in the model, we aim to control for the possibility that the dependent variable might influence the independent variable, thus mitigating concerns about reverse causality. This method has been applied in previous research on board gender diversity (Brodmann et al., 2022; Fu et al., 2023). Incorporating future values of the dependent variable enhances the assessment of the causal relationship between the variables and strengthens the validity of the findings. Consequently, the earlier regression model is extended to include these lead values as follows:

$$\begin{aligned}\text{Women on boards}_{it+1} = \\ \beta_0 + \beta_1 \text{Foreign institutional ownership}_{it} \\ + \beta_2 \text{Board size}_{it} + \beta_3 \text{Firm size}_{it} \\ + \beta_3 \text{Local institutional ownership}_{it} \\ + \alpha_i + \gamma_t + \epsilon_{it}\end{aligned}$$

As shown in Table 7, most of the regression coefficients are consistent with the results of the regression model without lead specifications. Using the lead model specification helps mitigate the potential for reverse causality by demon-

Table 7: Regression Results with Lead Model Specification

	Dependent Variables: Female Commissioners				Dependent Variables: Female Directors			
	$\%FC_{(t+1)}$		Dummy $FC_{(t+1)}$		$\%FD_{(t+1)}$		Dummy $FD_{(t+1)}$	
Explanatory Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$\%FII_{(t)}$	-0.103**	0.069	-0.119	0.232	0.113**	-0.035	0.241**	-0.128
	(0.049)	(0.064)	(0.113)	(0.197)	(0.044)	(0.029)	(0.114)	(0.097)
BOC $Size_{(t)}$	-0.002	-0.016	0.046***	-0.031				
	(0.004)	(0.015)	(0.010)	(0.036)				
BOD $Size_{(t)}$					-0.013**	0.003	0.032**	0.027
					(0.005)	(0.008)	(0.013)	(0.025)
Firm Size $(\ln)_{(t)}$	0.008***	-0.003	0.016**	-0.028	0.013***	0.016	0.026***	0.029
	(0.002)	(0.003)	(0.007)	(0.019)	(0.002)	(0.02)	(0.007)	(0.35)
$\%LII_{(t)}$	-0.163***	0.026	-0.272***	0.094	0.086**	-0.019	0.108	-0.083
	(0.046)	(0.047)	(0.105)	(0.120)	(0.043)	(0.031)	(0.107)	(0.0104)
N Observations	441	441	441	441	441	441	441	441
R-squared	0.04	0.903	0.07	0.867	0.051	0.942	0.083	0.889
Adjusted R2	0.05	0.852	0.06	0.796	0.042	0.911	0.074	0.831
Two-way Fixed Effects	No	Yes	No	Yes	No	Yes	No	Yes

This table presents regression results with robust standard errors using a one-period lead of the dependent variable, covering 441 observations. Columns are grouped by dependent variable: (1)-(2) percentage of female commissioners (%FC), (3)-(4) dummy for female commissioners, (5)-6) percentage of female directors (%FD), and (7)-(8) dummy for female directors. Dummy variables equal 1 if a firm has at least one female board member, 0 otherwise. All variables use an OLS regression model, except for models with dummy dependent variables, which use a linear probability model. The main independent variable is the percentage of foreign institutional investors (%FII), with controls for percentage of local institutional investors (%LII), board size, and firm size. Coefficients are shown with robust standard errors in brackets. Significance levels: (***) p<0.01, (**) p<0.05, (*) p<0.10.

strating that the current level of foreign and local institutional ownership influences future board gender diversity outcomes. After controlling for reverse causality, the results indicate that while institutional investors may drive changes in female directorships in subsequent years, the effects on female commissioners (BOC) are less pronounced. Columns (1) and (5) of Table 7 show that both types of investors are positively associated with female directors in the following year but negatively associated with female commissioners in the following year. This reinforces the idea that institutional investors, especially foreign ones, tend to focus more on increasing female representation in decision-making roles (the BOD) rather than monitoring positions (the BOC). However, these associations lose their significance when controlling for firm-specific and time-invariant factors.

In addition, board size and firm size are significantly associated with female member compositions in the following year. Of all the model specifications, only one negative association is observed between the BOD size and the proportion of female directors. Suggesting that a higher BOD size correlates with a lower percentage of female directors in the upcoming year. Meanwhile, other model specifications show that a larger board or firm size is positively linked to an increase in female member composition in the following year. Nevertheless, these associations are not robust to within-firm characteristics and temporal trends, as their significance diminishes when using two-way fixed effects.

Ultimately, using lead model specifications controls the possibility of reverse causality between the dependent and independent variables. By incorporating future values of the dependent variable, the lead model strengthens the validity of these findings and supports the argument that institutional investors contribute to shaping board gender diversity over time rather than simply responding to existing gender compositions. However, the robustness of these findings remains sensitive to firm-specific and temporal factors.

5. Conclusions

5.1. Research Conclusions

Board gender diversity is a critical aspect of corporate governance, with many stakeholders advocating for increased female representation in the boardroom. However, despite a global push for gender diversity, progress among Indonesian companies has been slow. This study reveals that

female representation in both the BOC and the BOD of Indonesian firms has rarely increased over time. While nearly half of the sample firms have at least one female board member, the proportion of women on boards has remained modest, staying below 20% from 2019 to 2022. Notably, female commissioners (BOC) are more likely to serve as the board's president than female directors (BOD), even though there are generally more female directors. Despite this, women remain underrepresented in key leadership roles, likely due to the limited pool of female candidates available.

My study emphasizes the ongoing underrepresentation of women in Indonesian boardrooms. After employing OLS and Linear Probability Regression models, foreign institutional investors have limited influence on driving board gender diversity in Indonesian public companies. Initially, foreign investors appeared to be positively associated with the number of female board members within the BOD, but this effect diminished upon closer examination using two-way fixed effects. Foreign institutional investors are negatively associated with the proportion of female directors, and similar negative associations were found between foreign investors and female representation in the BOC. These findings remain robust after addressing potential endogeneity concerns through lead model specifications.

As expected, local institutional investors show a consistent negative association with board gender diversity, suggesting that higher local ownership decreases the likelihood of a gender-diverse board. In contrast, board size and firm size are positively linked to the possibility of having at least one woman in both the BOC and BOD.

This research suggests that factors beyond those examined in this study may also influence female board representation in Indonesia. Future research could explore additional drivers, such as those identified by Abdullah (2014) and Shams (2019), who studied board gender diversity in Malaysia and other countries with similar characteristics. Variables such as firm ownership types, labor welfare regulations, and cultural factors may offer further insights into the underlying causes of female underrepresentation on boards in Indonesia.

5.2. Research Contributions and Limitations

The findings of my study contribute significantly to the academic literature on the drivers of board gender diversity, particularly in environments without formal mandates. Specifically, my research addresses key gaps in previous studies by exploring an under-researched context, focusing on Indonesia. This setting has rarely been examined in academic literature on board gender diversity. This approach provides valuable insights into an emerging market with unique governance structures, enriching the global understanding of the issue. Additionally, my research utilizes panel data incorporating Indonesian companies across multiple industry sectors, filling a gap in the existing literature, which often lacks a comprehensive view of board gender diversity in Indonesia. Furthermore, the study introduces a previously unexamined factor, foreign institutional ownership, into the analysis of board gender diversity, providing novel perspectives on potential drivers of change in Indonesian boardrooms. These contributions advance the literature and underscore the call for increased female representation in leadership positions within Indonesian public companies, particularly in the BOC and BOD. Ultimately, the study highlights the need for more direct approaches or policies to enhance gender diversity in the boardroom.

Despite these contributions, my study faces several limitations. First, the reliance on manually collected data has limited the sample size. Although efforts were made to ensure the representativeness of the sample, a larger sample could offer deeper insights into the influence of foreign ownership on board gender diversity. Expanding the dataset would provide a more nuanced analysis and more substantial conclusions. Second, while appropriate for this study's scope, using a reduced regression framework may not capture all complexities of the relationships between variables. This could lead to omitted variable bias, which may affect the internal validity of the results. Future research could address this by incorporating more complex models to examine broader factors influencing board gender diversity. Third, the geographic focus on Indonesian companies means that the findings may not be directly applicable to other contexts, limiting the external validity of the results. Comparative studies across different regions could help generalize the findings and provide a more comprehensive understanding of board gender diversity in varying contexts. Lastly, my analysis is confined to a specific period from 2019 to 2022, which may not fully capture evolving trends in board gender diversity. Future research could extend this timeframe to explore the long-term dynamics of gender diversity in boardrooms and consider broader data collection methods to enrich the understanding of these dynamics.

References

Abdullah, S. N. (2014). The causes of gender diversity in Malaysian large firms. *Journal of Management and Governance*, *18*(4), 1137–1159. https://doi.org/10.1007/s10997-013-9279-0

Adams, R. B. (2016). Women on boards: The superheroes of tomorrow? *Leadership Quarterly*, *27*(3), 371–386. https://doi.org/10.1016/j.leaqua.2015.11.001

Adams, R. B., & Ferreira, D. (2009). Women in the boardroom and their impact on governance and performance. *Journal of Financial Economics*, *94*(2), 291–309. https://doi.org/10.1016/j.jfineco.2008.10.007

Adams, R. B., & Funk, P. (2012). Beyond the glass ceiling: Does gender matter? *Management Science*, *58*(2), 219–235. https://doi.org/10.1287/mnsc.1110.1452

Admati, A. R., & Pfleiderer, P. (2009). The "Wall Street Walk" and shareholder activism: Exit as a Form of Voice. *Review of Financial Studies*, *22*(7), 2645–2685. https://doi.org/10.1093/rfs/hhp037

Bauer, R., Moers, F., & Viehs, M. (2015). Who withdraws shareholder proposals and does it matter? An analysis of sponsor identity and pay practices. *Corporate Governance: An International Review*, *23*(6), 472–488. https://doi.org/10.1111/corg.12109

BPS-Statistics Indonesia. (2023). Labor Force Situation in Indonesia.

Brodmann, J., Hossain, A., & Singhvi, M. (2022). Chief executive officer power and board gender diversity. *Finance Research Letters*, *44*. https://doi.org/10.1016/j.frl.2021.102099

Brown, K., Loriot, B., Peranginangin, Y., & Skully, M. T. (2023). Board Diversity and ESG: Australia, Indonesia, Malaysia and Thailand. https://ssrn.com/abstract=4625103

Csonka, A., & Milhomem, C. (2023). Women on Boards and Beyond: 2023 Progress Report.

Darmadi, S. (2011). Board diversity and firm performance: the Indonesian evidence. *Corporate Ownership and Control*, *8*. https://ssrn.com/abstract=1727195

Daromes, F. E., & Djie, M. (2019). Women on boards and greenhouse gas emission disclosures: How their impact on corporate reputation. https://ssrn.com/abstract=3434917

Deloitte Global. (2022). Women in the boardroom: A global perspective.

Deloitte Indonesia. (2021). Investment window into Indonesia. https://www2.deloitte.com/content/dam/Deloitte/id/Documents/about-deloitte/id-about-investment-window-indonesia-(IWI)-en-2021-2022.pdf

Fauver, L., Hung, M., Taboada, A. G., & Wang, E. J. (2022). Boardroom gender diversity reforms and institutional monitoring: global evidence. *Review of Accounting Studies*. https://doi.org/10.1007/s11142-022-09710-3

Fauziah, E. I., Probohudono, A. N., & Setiawan, D. (2022). The effect of board gender diversity on dividend payments: Evidence from Indonesia. *Australasian Accounting, Business and Finance Journal*, *16*(6). https://doi.org/10.14453/aabfj.v16i6.02

Financial Services Authority. (2023). Press release: The financial services sector remained resilient amidst continuing divergence in global economic recovery. https://ojk.go.id/en/berita-dan-kegiatan/siaran-pers/Pages/The-Financial-Services-Sector-Remained-Resilient-amidst-Continuing-Divergence-in-Global-Economic-Recovery.aspx

Fu, X., Li, Y., & Zhang, Z. (2023). Institutional investors, non-mandatory regulations, and board gender diversity. *Finance Research Letters*, *58*. https://doi.org/10.1016/j.frl.2023.104509

Goodman, J., Louche, C., van Cranenburgh, K. C., & Arenas, D. (2014). Social Shareholder Engagement: The Dynamics of Voice and Exit. *Journal of Business Ethics*, *125*(2), 193–210. https://doi.org/10.1007/s10551-013-1890-0

Gormley, T. A., Gupta, V. K., Matsa, D. A., Mortal, S. C., & Yang, L. (2023). The Big Three and board gender diversity: The effectiveness of shareholder voice. *Journal of Financial Economics*, *149*(2), 323–348. https://doi.org/10.1016/j.jfineco.2023.04.001

Hillman, A. J., Shropshire, C., & Cannella, A. A. (2007). Organizational predictors of women on corporate boards. *Academy of Management Journal*. http://www.jstor.org/stable/20159898

IMF. (2023). World Economic Outlook Database. https://www.imf.org/en/Publications/WEO/weo-database/2023/October

ISS. (n.d.). ISS Global Voting Principles. https://www.issgovernance.com/policy-gateway/iss-global-voting-principles/

Kang, E., Ding, D. K., & Charoenwong, C. (2010). Investor reaction to women directors. *Journal of Business Research*, *63*(8), 888–894. https://doi.org/10.1016/j.jbusres.2009.06.008

Kirsch, A. (2018). The gender composition of corporate boards: A review and research agenda. *Leadership Quarterly*, *29*(2), 346–364. https://doi.org/10.1016/j.leaqua.2017.06.001

Matanda, T., Wang, C., & Emelianova, O. (2023). Women on Boards: Progress Report 2022.

NSWI. (2023). Foreign direct investment realization by country. https://nswi.bkpm.go.id/data_statistik

Perrault, E. (2015). Why Does Board Gender Diversity Matter and How Do We Get There? The Role of Shareholder Activism in Deinstitutionalizing Old Boys' Networks. *Journal of Business Ethics*, *128*(1), 149–165. https://doi.org/10.1007/s10551-014-2092-0

PRI. (2023). A guide to filing impactful shareholder proposals.

Quek, H. (2024). Women directorships highest on record across three sectors in 2023. *The Business Time Singapore*. https://www.businesstimes.com.sg/companies-markets/women-directorships-highest-record-across-three-sectors-2023

Rastad, M., & Dobson, J. (2022). Gender diversity on corporate boards: Evaluating the effectiveness of shareholder activism. *Quarterly Review of Economics and Finance*, *84*, 446–461. https://doi.org/10.1016/j.qref.2020.09.007

Refinitiv. (2022). ENVIRONMENTAL, SOCIAL AND GOVERNANCE SCORES FROM REFINITIV - May 2022. https://www.refinitiv.com/content/dam/marketing/en_us/documents/methodology/refinitiv-esg-scores-methodology.pdf

Rojas, M., M'Zali, B., Turcotte, M. F., & Merrigan, P. (2009). Bringing about changes to corporate social policy through shareholder activism: Filers, issues, targets, and success. *Business and Society Review*, *114*(2), 217–252. https://doi.org/10.1111/j.1467-8594.2009.00341.x

Ryan, L. V., & Schneider, M. (2002). The antecedents of institutional investor activism. *Academy of Management Review*, *27*(4), 554–573. https://doi.org/10.2307/4134403

Selby, C. C. (2000). From Male Locker Room to Co-Ed Board Room: A Twenty-Five Year Perspective.

Shams, A. (2019). *Women on corporate boards in Asia: The influence of formal and informal institutional features* [Order No. 28758841]. https://www.proquest.com/dissertations-theses/women-on-corporate-boards-asia-influence-formal/docview/2604439504/se-2

Storvik, A., & Teigen, M. (2010). Women on board - the Norwegian experience. https://library.fes.de/pdf-files/id/ipa/07309.pdf

Tang, H., & Zhang, Y. (2021). Do multinationals transfer culture? Evidence on female employment in China. *Journal of International Economics*, *133*. https://doi.org/10.1016/j.jinteco.2021.103518

Terjesen, S., Aguilera, R. V., & Lorenz, R. (2015). Legislating a woman's seat on the board: Institutional factors driving gender quotas for boards of directors. *Journal of Business Ethics*, *128*(2), 233–251. https://doi.org/10.1007/s10551-014-2083-1

The Indonesian Ministry of Labor. (2023). Workforce Data. https://satudata.kemnaker.go.id/publikasi/140

The Indonesian Ministry of Women Empowerment and Child Protection. (2022). Indonesian Women Profile. https://www.kemenpppa.go.id/page/view/NDM3NA==

Tjondro, E., Kristiany, S., & Sanjaya, C. N. (2020). Women on the Executive Board and Woman CEO: Indonesia's Financial Firm. *Advances in Economics, Business and Management Research*, *158*. https://doi.org/10.2991/aebmr.k.201212.053

Junior Management Science 10(3) (2025) 631-656



Junior Management Science

www.jums.academy
ISSN: 2942-1861



Value Creation Opportunities of Generative AI – A Case Study

Alexander Sake

Technical University of Munich

Abstract

The transformative potential of Generative AI promises novel capabilities within business environments. This study examines the value creation potential of Generative AI within a large multinational corporation. A single case study approach at Siemens was employed, combining extensive observations, interviews, and the application of existing AI frameworks. Findings reveal diverse use cases demonstrating value creation potential, particularly through smart assistants and lighthouse projects. This thesis proposes a novel framework for Generative AI adoption, emphasizing the distinctive exploration phase made possible by the technology's accessibility to non-technical domain experts, while also outlining essential scaling strategies. This study offers valuable insights into a company's approach to Generative AI, provides practical implications, and expands ongoing research on AI-driven value creation.

Keywords: artificial intelligence; exploration; generative AI; scaling; technology adoption; use cases; value creation

1. Introduction

The release of ChatGPT in November 2022 marked a significant moment where the public could, for the first time, directly engage with the latest advances in artificial intelligence (AI). Since its launch, the tool's impact and capabilities have been a topic of diverse discussions among journalists, scientists, managers, and governments, oscillating between praise and cautionary notes. This thesis intends to shed light on the latest advances of the technology, to describe the value creation opportunities that arise from it and to help companies identifying requisites and capabilities for a successful introduction into their organization.

Many studies have been conducted on the impact of AI on companies and society. Following technological advancements, these studies often predicted new value creation opportunities arising from the successful implementation of AI solutions in organizations. Missing out on AI, results frequently indicated a significant gap in future performance and competitiveness. A study highlighting these aspects, was conducted by Accenture (Reilly et al., 2019). Based on a global survey among 1,500 C-suite executives, they discovered that 84% of leaders believed to achieve their growth ambitions only with the help of AI. At the same time, 76% struggle to scale the technology across their organization.

The rise of Generative AI was accompanied by many studies as well, evaluating real-world use cases and analyzing the impact on organizational processes and the future of work. In particular, when the GPT-4 model was released, questions arose how the tool will transform and potentially replace activities of knowledge workers (Dwivedi et al., 2023, p. 7). A study by BCG among 750 consultants revealed a performance increase of 40% when using GPT-4 for typical consulting activities compared to a control group (Dell'Acqua et al., 2023, p. 17). This was accompanied by an increase in speed of 25% with a positive impact across all skill levels. Focusing on the performance of customer support agents, Brynjolfsson et al. (2023) also examined the impact of Genera-

I would like to express my sincere gratitude to my supervisor, Tim Lämmermann, M.Sc., for his valuable support and guidance throughout my master thesis. I also extend my special thanks to Prof. Dr. Thomas Hutzschenreuter, Chair of Strategic and International Management, for his inspiring teaching and expertise. I am deeply grateful to all the interview partners who contributed to my work with their valuable insights and experiences. A big thank you goes to my managers for allowing me to pursue this project alongside my job. Lastly, I would like to thank my family and friends – your understanding and support during the intense period of the master thesis were invaluable.

DOI: https://doi.org/10.5282/jums/v10i3pp631-656

© The Author(s) 2025. Published by Junior Management Science.
This is an Open Access article distributed under the terms of the CC-BY-4.0 (Attribution 4.0 International). Open Access funding provided by ZBW.

tive AI. Based on data from 5,176 agents, they found that AI was able to increase productivity by 14%. In their survey, differences in regard to skill-levels became apparent: While new or low-skilled workers' performance increased by 34%, highly skilled workers barely benefited from the tool (2023, p. 1). While these examples show the potential added value in business settings, other studies focus on the impact on job roles and the global economy. Findings by Goldman Sachs Research suggest that Generative AI could have a "profound" effect on the world economy and society, potentially raising global GDP by 7% over a 10-year period (Goldman Sachs, 2023). A report by Eloundou et al. (2023) focusing on the US labor market see 80% of the workforce impacted the Generative AI by at least 10% while for almost 20% of the workforce, half of the tasks are seen to be exposed to AI (2023, p. 1).

2. Research Objective

Generative AI represents a significant advancement in artificial intelligence, characterized by its ability to produce new content through pattern recognition in existing data (Feuerriegel et al., 2023, p. 1). This is a shift from traditional AI, which primarily dealt with tasks such as identification and classification (LeCun et al., 2015, p. 436). A major development in Generative AI was the introduction of the Transformer architecture in 2017 (Vaswani et al., 2017), which enhanced the AI's capability to comprehend and process extensive information, thereby improving its content generation ability. Further advancements were achieved with the introduction of the Generative Pre-trained Transformer (GPT) model by Radford and Narasimhan (2018), representing a significant step forward in the field of AI-driven text generation. Trained on diverse datasets, the model was capable of producing text that is not only coherent but also contextually relevant.

The primary **Research Question (RQ1)** intends to delve into the advancements introduced by Vaswani et al. (2017) and Radford and Narasimhan (2018). It aims to investigate how Generative AI diverges from previous AI models in terms of learning, information processing, and content creation. Grasping these attributes of Generative AI is key to understanding its transformative impact. To address these considerations, the research question is formulated as follows:

> *RQ1: How does the architecture and functionality of Generative AI differ from previous AI models?*

The outlined studies conducted by Brynjolfsson et al. (2023) and Dell'Acqua et al. (2023) provide initial insights into the potential of Generative AI to enhance workforce productivity. However, further research is necessary to gain a deeper understanding of the technology and its potential benefits, particularly for larger organizations. The disruptive nature of AI has long been a subject of investigation, yet many companies struggle to translate AI's potential into tangible business value Shollo et al., 2022, p. 1 To address this gap, it is crucial to analyze the characteristics of Generative AI and explore how it can benefit larger corporations on a broad scale. This exploration is essential for identifying and understanding the opportunities that Generative AI presents for creating value within large corporate environments. By doing so, the thesis aims to contribute to the ongoing discussion on the value creation potential of AI. Expressed in **Research Question 2:**

> *RQ2: To what extent does Generative AI open up new opportunities for value creation in companies?*

Based on the potential value-add, the question arises how companies can organize and drive the implementation of AI in their organization. Previous research has shown that multiple dimensions need to be considered to adopt AI successfully (Uren & Edwards, 2023). However, depending on the focus area of the research, different requirements have been proposed. While some studies focus on technological aspects of AI (Davenport & Ronanki, 2018, p. 52), others highlight the importance of strategic considerations (Brock & Wangenheim, 2019, p. 7). Providing insights from a case study could help to verify the various aspects of these studies, strengthen the outlined models, and add to the ongoing discourse in the field of science. Articulated in **Research Question 3**:

> *RQ3: What considerations should companies make in order to exploit the value creation potential of Generative AI?*

The thesis aims to provide valuable insights for both scholars and professionals through exploring the defined research questions. Firstly, it seeks to evaluate existing models in real-world scenarios, helping to understand their practical use and how earlier research relates to Generative AI. Secondly, it aims to offer practitioners useful insights into potential uses and strategies for scaling Generative AI in their organizations.

In concluding this chapter, it is vital to highlight that a comprehensive interpretation of the term Generative AI should be maintained throughout the thesis. Given the exploratory character of the case study and the novelty of the technology, in the interviews, a clear distinction between different technological aspects was not always made. This holds in particular true for the terms "Generative AI", "ChatGPT" and "Large Language Model (LLM)", which have been frequently used. As this work is about identifying value creation opportunities from a business perspective rather than detailing exact technical mechanisms, this should be considered an acceptable inaccuracy.

3. Conceptual and Theoretical Background

The next chapter delves into the conceptual and theoretical framework essential for the thesis. It begins by highlighting key developmental stages that have led to Generative AI. Following that, it introduces theoretical concepts that shed light on how AI can enhance value creation and the essential capabilities required by companies to achieve this.

3.1. Evolution of Artificial Intelligence and Machine Learning

The idea of artificial intelligence dates back to the 1950s when the mathematician and computer scientist Alan Touring asked himself how the intelligence of computers could be measured (Turing, 1950). Instead of asking whether a machine is intelligent, he introduced the "Imitation Game" (now known as the Turing Test) to ask whether machines can imitate human responses convincingly (Turing, 1950, p. 433). Turing described machines or digital computers as complex systems capable of a wide range of tasks, similar to a human following instructions (Turing, 1950, p. 436). He suggested that these computers, due to their vast capabilities, could potentially pass as human in his test (Turing, 1950, p. 442). This idea was a significant step in understanding machine intelligence, proposing a practical way to measure it and laying the groundwork for the field of artificial intelligence.

The term "Artificial Intelligence" itself only became a collective term for a variety of different concepts in 1956. McCarthy, Assistant Professor of Mathematics at Dartmouth College, chose the term for a workshop on the topic and is therefore regarded today, alongside Marvin Minsky, Allen Newell, Herbert Simon as one of the "fathers of AI" (Nilsson, 2013, p. 80). McCarthy was among the first who outlined the requirements for a system to evolve human-like intelligence in more detail. In his paper "'programs with common sense" (1958), he discussed features which would be required for a machine to evolve intelligence. These included the capability to represent all behaviors, simply express interesting behavioral changes, improve most aspects of behavior, understand partial success in complex problems, and create improvable subroutines (McCarthy, 1958, p. 5).

Building on McCarthy's foundational work, various definitions of AI have emerged over time. Simmons and Chappell (1988, p. 14) define AI as the behavior of a machine which, if a human behaves in the same way, is considered intelligent. Luger and Stubblefield (1998, p. 1) define AI as a branch of computer science focused on the automation of intelligent behavior. In contrast, Russell and Norvig (2021, p. 2) follow an rational agent approach, defining AI as the study and construction of intelligent agents. Based on the scientific work, the following definition shall be applied in this thesis: AI is defined as the ability of machines or computers to learn and perform tasks that are typically attributed to human intelligence.

The idea of teaching machines to learn can be traced back to Samuel (1959). Using the game of checkers, he was able to program a computer that was able to play the game and outperform a human player based on a rudimentary set of parameters and rules (Samuel, 1959, p. 535). His novel concept of machines that could improve their performance over time through experience, is often cited as one of the earliest work of "machine learning" (McCarthy & Feigenbaum, 1990, p. 10).

What does "learning" actually mean when discussing computer programs? In his book "Machine Learning" (1997), T. Mitchell offers the following definition: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E" (Mitchell, 1997, p. 2). He also uses the game of checkers as an example to describe a computer program that improves its ability to win by gaining experience. This example can be applied to more than simple games. More complex rules and sequences of instructions, better known as algorithms, now enable computer programs to learn a wide variety of tasks. Speech recognition, driving autonomous cars and the classification of new astronomical structures are examples where machine learning is used in practice today (Mitchell, 1997, p. 3).

Before delving into the specifics of machine learning, **Figure 1** presents key terms that are beneficial for subsequent exploration of the topic. Machine learning, which can be seen as a subfield of artificial intelligence, encompasses diverse learning paradigms. These will be outlined in the following section, preceding in-depth explorations of deep learning and Generative AI in later chapters.

The first type of learning is called supervised learning (Goodfellow et al., 2016, p. 103). In this method, the machine or algorithm is provided with a dataset whose examples are labeled. When new, unlabeled data is presented, the machine attempts to predict the label based on the patterns and properties of the known data. The result is then compared with the actual label. The training of the algorithm consists of reducing the error between the estimated and the actual labels, enabling the machine to recognize new input and classify it correctly (LeCun et al., 2015, p. 436). The second type is called self- or unsupervised learning. Instead of providing the machine with labeled data, it learns to identify patterns and properties from the data itself. This way, it learns about the probability distribution of the entire dataset (Goodfellow et al., 2016, p. 103). While this type of learning has been explored for a long time (see Hinton and Sejnowski, 1985), it gained popularity with the introduction of Transformer architecture, large datasets, and the availability of computing power (Radford et al., 2019, p. 10). A third type is called reinforcement learning, which introduces a feedback loop into the learning process. During training, the model uses its experiences to improve performance through a reward and penalty system, aiming to increase its cumulative rewards over time (Sutton & Barto, 2018, p. 2).

3.1.1. From Turing to Deep Learning

Over the past few decades, research in the field of AI has made substantial progress. New methodologies, models, and architectures have been developed, vastly surpassing initial conceptions. These advancements enable AI to perform tasks such as writing poems, developing software, and composing music (Feuerriegel et al., 2023, p. 1). A pivotal technique that has been central to these capabilities is deep learning (Goodfellow et al., 2016, p. 5), which involves decomposing complex relationships into simpler, interconnected concepts. These simpler concepts can then be represented by even more

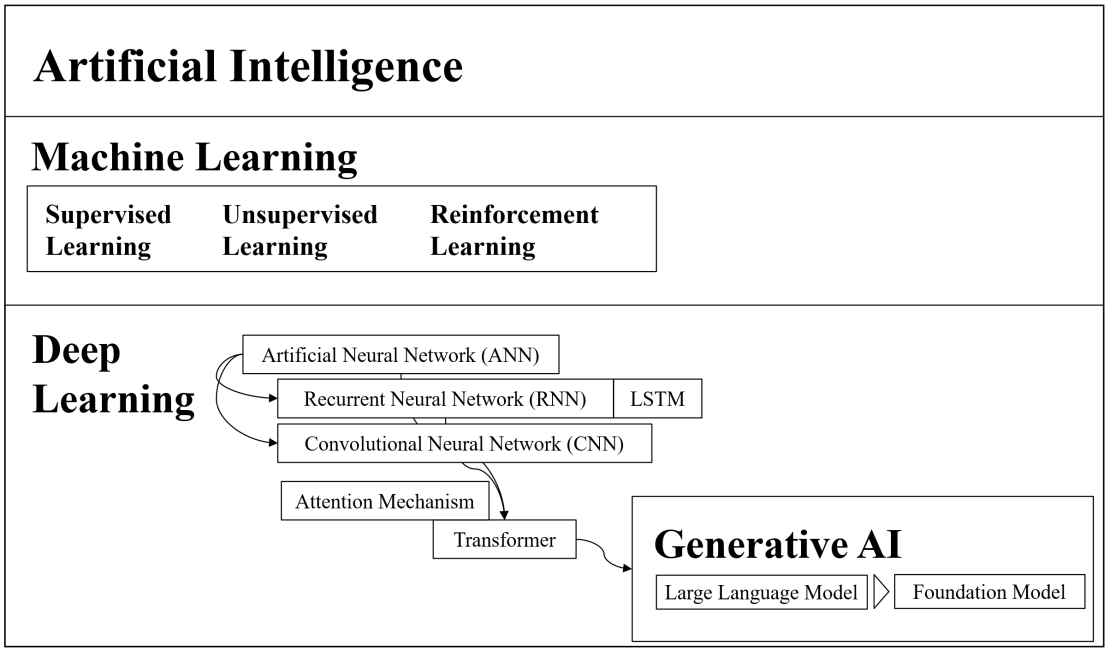


Figure 1: Relationship between AI terms which will be discussed in this thesis.

fundamental concepts (Goodfellow et al., 2016, p. 1). In this way, deep learning enables computers to infer complex relationships from simpler ones (Goodfellow et al., 2016, p. 2). The architecture of these models comprises numerous hierarchical layers, which is why it is referred to as "deep learning".

Deep learning is based on the concept of artificial neural networks (ANN), that is, systems inspired by the human brain's structure (Goodfellow et al., 2016, p. 165). They are termed networks because they typically comprise a large number of interconnected nodes, referred to as neurons, distributed across various layers (Goodfellow et al., 2016, p. 164). **Figure 2** illustrates the simplified structure of an ANN.

The input layer receives the initial information (for example, raw data), which is then transmitted to the subsequent layer. Each node or neuron processes this input through a simple computation and forwards its output to the next layer of neurons. The artificial network "learns" by modifying the parameters of the connections between individual neurons, known as weights, based on experience and performance. This adjustment is made using a function that calculates the discrepancy between the actual output and the desired output, then alters the weights to minimize this error (LeCun et al., 2015, p. 436) This process, known as back-propagation, was first detailed in an article by Rumelhart et al. (1986, p. 533) and enables the program to self-organize and refine its internal structure. The layers between the input and output are called "hidden layers" because their activities are not directly observable from the input data. Instead, the model must determine which patterns are significant for explaining the input data (Goodfellow et al., 2016, p. 6).

There are different types of artificial neural networks. Two of these are recurrent neural networks (RNN) and convolutional neural networks (CNN). While RNNs specialize in the processing of sequences, such as text (Graves, 2012, p. 1), the strength of CNNs lies in the processing of raster information, such as images (Goodfellow et al., 2016, p. 367). A weak point of the RNN and CNN architectures is the limited context window, i.e., the amount of information that the system can store over a longer period of time (Hochreiter et al., 2001, p. 11). To mitigate this vulnerability, Hochreiter and Schmidhuber (1997) developed a novel RNN architecture called Long Short-Term Memory (LSTM). Their customization made it possible to solve larger and more complex tasks that were not feasible with the standard RNN architecture (Hochreiter & Schmidhuber, 1997, p. 2).

However, the processing of long sequences remained a major challenge. This became particularly apparent in the area of machine translation, as it was difficult for RNNs to recognize word dependencies over long distances in sentences or paragraphs (Bahdanau et al., 2014, p. 6; Kim et al., 2017, p. 2). One idea was not to encode the entire sequence, but rather to focus on individual sections. The so-called "attention mechanism" allows the model to dynamically focus on different parts of the input (Kim et al., 2017, p. 2). To simplify, this can be compared to reading a scientific article and stumbling over a difficult section. Instead of trying to comprehend the entire content, it is helpful to focus on individual words or to look back at the previous sections. In this way, attention is shifted to the key aspects, helping to break down the section or sequence into smaller, more manageable parts and to understand how each part connects to the others.

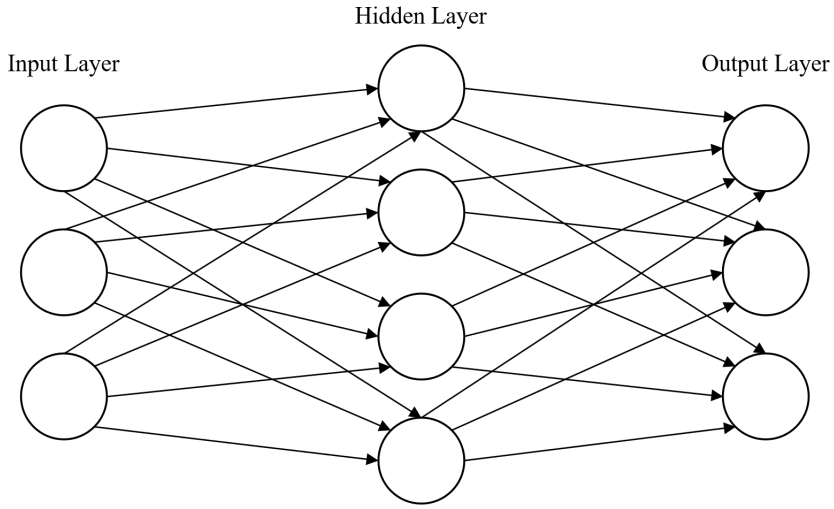


Figure 2: Simplified structure of an Artificial Neural Network (ANN)

3.1.2. Emergence of Attention Mechanisms and Transformers

The attention mechanism described above was very effective, especially for machine translation tasks, and is therefore regularly used in deep learning models (Kim et al., 2017, p. 1). However, the sequential nature of RNNs remained the limiting factor in terms of parallelization and scalability of these models (Peng et al., 2023, p. 1). In 2017, Vaswani et al. proposed a new architecture in their article "Attention Is All You Need". Called the "Transformer", this new model is based solely on the attention mechanism and no longer on an RNN or CNN structure. Its strength lies in its ability to reference infinitely long sequences without the previous limitations of RNNs (Vaswani et al., 2017, p. 5), as visualized in **Figure 3**.

If the model is given the task to answer a question, it can refer to the entire text with the help of the attention mechanism. This is particularly helpful when textual relationships only become apparent upon viewing an entire paragraph. RNNs, with their limited context window, were restricted in this regard. In the example, an RNN model would "forget" that the question referred to a cat.

In 2015, the research group OpenAI was formed with the goal of advancing in the field of AI (OpenAI, 2015). One of the first articles published by OpenAI's researchers related to the Transformer architecture was "Improving Language Understanding by Generative Pre-Training" (2018). In the article, Alec Radford and Karthik Narasimhan focus on enhancing natural language processing (NLP) via a semi-supervised learning approach that combines unsupervised pre-training and supervised fine-tuning (2018, p. 1). The authors propose a novel approach that utilizes large unlabeled text corpora for pre-training a language model, followed by task-specific fine-tuning. Their approach aims to overcome the limitations of supervised models, which often require extensive labeled data that is scarce or expensive to obtain.

In their model, Radford and Narasimhan (2018, p. 2) employ the Transformer architecture due to its efficient handling of long-term dependencies in text sequences. Their training method consists of two stages: In the pre-training phase, they use a language modeling objective on an unsupervised corpus, establishing the initial parameters of the model. Language modeling is the process of training a model to predict the next word based on previous words in a sentence (Voita et al., 2019, p. 3). **Figure 4** shows a simplified example using a Google search, where the model suggests the next word based on its predicted probability.

Applying this principle to training data, a set of 7,000 unpublished books, the model was able to recognize correlations and gain knowledge not only about individual sentences but also about the nature of language in general (Radford & Narasimhan, 2018, p. 8). In the fine-tuning stage, the learned correlations and knowledge were adapted to specific language understanding tasks such as question answering, common sense reasoning, semantic similarity analysis, and text classification. Subsequently, they used different tests to measure the performance in these categories, achieving state-of-the-art results in 9 of 12 categories (Radford & Narasimhan, 2018, p. 8).

3.1.3. Advancements in GPTs and Foundation Models

Radford and Narasimhan's (2018) work laid the basis for further development of the GPT architecture. In an article published in 2019, a next-generation GPT model, GPT-2, was introduced (Radford et al., 2019). Instead of using curated datasets, they based their training on publicly available data from the internet. With the intention of creating a large and diverse corpus of natural language text covering as many domains as possible, they gathered over 8 million documents, totaling 40 GB of data (Radford et al., 2019, p. 3). When testing for natural language processing tasks, such as question answering, translations, reading comprehension, or summarization, they observed the model's capability to learn these tasks without explicit supervision (Radford et al.,

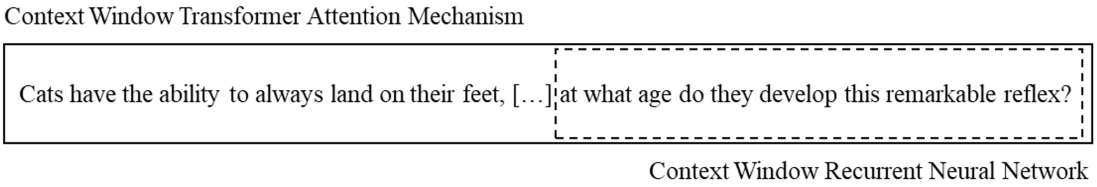


Figure 3: Comparison of context windows

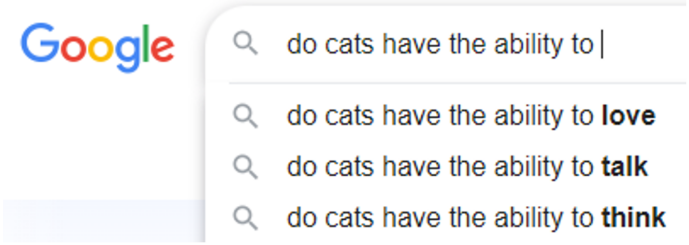


Figure 4: Google Search: Example of word predictons

2019, p. 1). This can be seen as an significant breakthrough, as previous models required substantial amounts of labeled and curated data to learn effectively (Radford & Narasimhan, 2018, p. 1). During their tests, they experimented with different model sizes and showed that performance improved with increasing model capacity. Their largest model, with 1.5 billion parameters, achieved state-of-the-art results in multiple settings (Radford et al., 2019, p. 10).

The potential for performance gains through increased model capacity was further explored by Brown et al. (2020). Their article introduced the GPT-3 model, containing 175 billion parameters and representing a significant advancement over previous models by a factor of 10. Confirming the researchers' hypotheses, the model outperformed its predecessor, GPT-2, and even rivaled the performance of state-of-the-art fine-tuned systems (Brown et al., 2020, p. 9). They found that additional computational power directly correlated with increased performance, laying the groundwork for further research. Moreover, since their work focused on creating a task-agnostic model, they anticipated that future fine-tuning would further enhance their model's performance (Brown et al., 2020, p. 2).

Further developments occurred over the next years, advancing the model and introducing fine-tuning to boost performance (Ouyang et al., 2022, p. 1). This culminated in OpenAI's release of the now well-known ChatGPT (OpenAI, 2022). With its unexpected capability to generate text indistinguishable from human writing and engage in believable human-like conversations, it caught many researchers off guard (Dwivedi et al., 2023, p. 4). Technologically, the advances stem from two primary aspects. Firstly, an updated GPT-3 model with a modified dataset serves as the foundation for ChatGPT (GPT-3.5). Secondly, a technique called Reinforcement Learning from Human Feedback (RLHF) was used to fine-tune the model. RLHF can be defined as a method where the model is improved based on human evaluations of its outputs, guiding it toward desired behaviors and responses (Ziegler et al., 2019, p. 1). The process utilizes a reward model that is trained using feedback from human labelers, who evaluate and rank the model's answers from best to worst. This reward model is then used within the main model, employing reinforcement learning for self-optimization. This significantly enhances the model's ability to provide answers aligned with human preferences and consider previous input in conversational settings (Feuerriegel et al., 2023, p. 4). Alongside the technological leap, OpenAI's decision to grant unrestricted public access to ChatGPT and its simplified user interface fueled its explosive worldwide adoption (Feuerriegel et al., 2023, p. 5). **Figure 5** shows the landing page of ChatGPT 3.5, offering examples of questions and a prompt bar.

In March 2023, just five months after ChatGPT's release, OpenAI unveiled the next generation of their GPT model (OpenAI, 2023). Unlike previous models focused solely on textual input, GPT-4 could process both text and images as input, generating textual outputs. While detailed specifications remain limited, its performance exceeds that of its predecessor, GPT-3.5, in numerous areas, achieving human-level results on multiple benchmarks (OpenAI, 2023, p. 6). As prior research demonstrated a correlation between performance and computational power, anonymous reports suggest the model may possess 1.76 trillion parameters (Bastian, 2023). While this information should be treated with caution, it gives an indication of the rapid technological development in the area of Generative AI. From 1.5 billion parameters in 2019, over 175 billion in 2020 to potentially 1.76 trillion in 2023.

With the advances in the field of AI, a shift in the development of AI models has emerged. While previous models were trained on labeled data and fine-tuned to specific tasks, newer models such as GPT-3 and 4 differ significantly in two areas. Firstly, instead of labeled data, they are trained

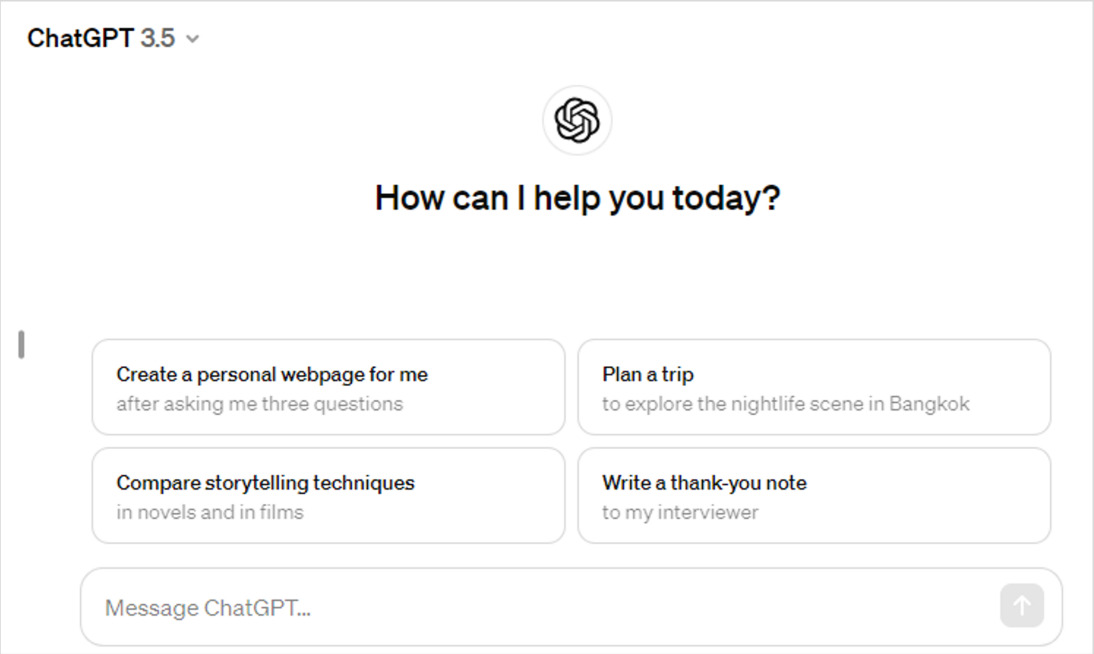


Figure 5: ChatGPT Landing Page.

on broad, unlabeled information using self-supervised learning. Secondly, the sheer scale allows the models to perform tasks they have never been explicitly trained on (Brown et al., 2020, p. 9). This general understanding of language and other patterns, fostered by the increase in available computational power, led to the need to create a new class for these models. In their scientific article, Bommasani et al. (2021) therefore proposed the term "foundation model" to encapsulate these characteristics and describe the emergence and application of these models in detail. According to the authors, foundation models build on the idea of large language models, which gain an understanding of natural language based on textual datasets (Brown et al., 2020, p. 9). Expanding on this concept, foundation models are multimodal – besides text, they are able to process data such as images, videos, 3D signals, and others. As with large language models, their vast datasets allow them to gain a general understanding of the underlying structures and properties applicable to different downstream tasks (Bommasani et al., 2021, p. 6). A prominent example is GPT-4, with its capability to provide answers based not only on text but also on images (OpenAI, 2023).

3.1.4. Generative AI: Scope and Capabilities

In the previous section, various development steps and technological innovations have been outlined, from simple ML algorithms to deep learning techniques, culminating in GPTs, large language and foundation models. One feature that was mentioned in connection with GPTs is their ability to generate new content. While the concept of GPTs has its origin in the publicly available work by Vaswani et al. (2017) on the Transformer architecture, the term itself was primarily coined by OpenAI's advances in the field (in particular with GPT-3 and 4). Google, Meta, and other companies have since published their own models based on the Transformer model (for example Google's Gemini or Meta's Llama 2). To maintain a neutral perspective, the term "Generative AI" is therefore used throughout this thesis.

Generative AI, from a technical standpoint, is rooted in generative modeling, which seeks to understand the joint probability distribution $P(X, Y)$ where X represents the data and Y denotes the labels. The objective is to grasp how data is generated in order to create new data points (Ng & Jordan, 2001, p. 1). This approach differs from discriminative modeling, which focuses on modeling the conditional probability $P(X, Y)$, representing the probability of the label Y given the input data X (Ng & Jordan, 2001, p. 1). This differentiation is crucial from an application standpoint: Generative AI models enable the generation of new data instances based on observed probability distributions within a given dataset (Goodfellow et al., 2016, p. 716), whereas discriminative modeling exhibits superior performance in classification tasks (Ng & Jordan, 2001, p. 1).

3.2. AI Value Creation and Business Impact

Over the last few decades, a large number of scientific articles have focused on the influence of AI on companies. The question of how and in which areas AI can be used and what prerequisites companies need to fulfil in order to do this successfully is a recurring one. Selected theoretical concepts will therefore be presented and described in the following section. Caner and Bhatti's (2020, p. 182) conceptual framework can be used as a starting point. Based on an extensive literature review, the authors propose six perspectives from which the strategic dimension of AI can be viewed:

1. Capabilities and Limitations of AI
2. Business Functions and AI
3. Tasks, Jobs, and Intelligence
4. Economy and AI
5. AI and Law, Regulations, Governance
6. Industries and AI

This thesis focuses on the first three aspects, as they are especially useful for understanding how AI impacts individual companies. The first aspect covers the **capabilities and limitations of AI**, as discussed in Chapter 3.1. It explores the technical aspects of AI, including its various applications, technologies, and limitations. Understanding AI's strengths and weaknesses is crucial in a business context to set realistic expectations and identify practical deployment options (Davenport & Ronanki, 2018, p. 110). While some constraints identified by Caner and Bhatti (2020), like data labeling and learning generalizability, have been addressed with technological progress (refer to OpenAI, 2023), challenges such as biases and lack of explainability persist.

The aim of the second dimension **business functions and AI** is to categorize the fields of application of AI. Caner and Bhatti base this on an IBM survey and a study by Davenport and Ronanki (2018) (Caner & Bhatti, 2020, p. 186). They differentiate between two subject areas: Firstly, the areas in which AI is used, e.g., in internal processes or customer services. Secondly, the way in which AI is used, e.g., to automate entire processes or gain new insights from data.

The third dimension, **tasks, jobs, and intelligence,** describes the different stages in which AI can be used. Based on Rao's (2017) framework, a distinction is made between three levels. Firstly, "Assisted Intelligence" includes all processes that are conducted in a conventional manner and are supported by AI. Secondly, "Augmented Intelligence" refers to areas where AI takes over a large part of the value creation (the areas mentioned are, for example, automatic translation and automatic analysis of legal documents). Thirdly, "Autonomous Intelligence" comprises those processes that can be conducted completely without human intervention in the future.

3.2.1. AI in Strategic Business Context

The progress that has been made in the field of artificial intelligence repeatedly raises the question of how this technology can be strategically implemented and utilized. In particular, the question of the sources of value creation is being researched in depth (Borges et al., 2021; Kitsios & Kamariotou, 2021; Trunk et al., 2020). This aligns with the thesis's objective, namely, to investigate the extent to which the existing concepts can be applied to Generative AI. Borges et al. (2021) extensive literature review provides the first insights into this area. Based on 41 studies, they define four source of value creation from AI:

1. Decision Support
2. Customer and Employee Engagement
3. Automation
4. New products and services

Decision support refers to the ability of AI to support humans in strategic and operational business decisions (Borges et al., 2021, p. 11). This occurs, for example, when deep learning techniques are used to detect patterns in data, subsequently guiding the decision-making processes. Consequently, decisions can be executed more swiftly and with greater reliability (Borges et al., 2021, p. 12). However, scientists highlight the necessity for further research in this rapidly evolving field, particularly concerning the interaction between humans and AI and its impact on organizational performance (Lichtenthaler, 2019, p. 8). This thesis presents an opportunity to examine these dynamics, especially in the context of Generative AI. It will be insightful to determine whether and how the technology can optimize decision-making processes and thereby contribute to value creation.

The second identified source of value creation is called **customer and employee engagement**. The aim is to use AI to improve the customer experience and internally to attract employees to the new technology. Although the article analyses various academic papers, the authors note that further research is needed as some examples cannot be generalized. Generative AI, with the ability to create content that is indistinguishable from human input (Feuerriegel et al., 2023, p. 1), could offer new opportunities in this context.

Automation represents the next source of value creation, allowing costs to be reduced and efficiency to be increased through the strategic use of AI technologies. The authors argue that a competitive advantage can also be created if processes can be automated more quickly than is possible for competitors (Borges et al., 2021, p. 12).

The fourth source of value creation **new products and services** deals with the ability to generate new business ideas through AI. Borges et al. (2021) suggest AI's potential in driving innovation and creating new products and services, however they could only identify and provide limited empirical evidence. Consequently, the authors argue that further research is needed to understand how AI can be used strategically to create new products and solutions (Borges et al., 2021, p. 12).

The sources of value creation listed by Caner and Bhatti as well as Borges et al. can also be found in similar form in other literature reviews and scientific articles. For example, in their literature review of 81 articles in the context of AI and business strategy, Kitsios and Kamariotou (2021) identified "AI and Machine Learning in organizations", "AI, knowledge management and decision-making" and "AI, service innovation and value" as sources of value creation.

An area that was not mentioned in earlier sources and which they discovered was the **alignment of AI tools and Information Technology (IT) with organizational strategy**.

According to the authors, the company's IT strategy should be closely aligned with the business strategy. They suggest developing a separate digital (business) strategy to take account of the increasing importance of IT (Kitsios & Kamariotou, 2021, p. 6). With regard to AI, the authors note that it can now perform cognitive tasks, i.e., processing new information, recalling it from memory and using it in communication with humans. One obstacle to the implementation of these technologies however is not the technology itself, but human perceptions, company processes and cultural barriers (Kitsios & Kamariotou, 2021, p. 7).

3.2.2. AI and Business Capabilities

When introducing AI, the question regularly arises as to what capabilities companies need. These requirements have been analyzed in detail in various scientific articles, each with a different focus area. Based on a comprehensive literature review, Reim et al. (2020) approach the topic from a business model perspective. They derive a four step model to provide guidance to companies (Reim et al., 2020, p. 180): In their view, companies need to develop a good understanding of AI, have a thorough understanding of their business model and possible areas for innovation, must invest in the needed capabilities to support AI implementation and aim for a high level of organizational acceptance. From their perspective, the initial steps will likely be driven by the upper management, laying the conceptual foundation for the introduction of AI and analyzing the required capabilities (Reim et al., 2020, p. 187).

Based on previous research on business process thinking and knowledge management, Uren and Edwards (2023) propose the following categories: People, Processes, Technology and Data (Uren & Edwards, 2023, p. 7). In their view, data in particular plays a crucial role as it provides the foundation for the successful adoption of AI (Uren & Edwards, 2023, p. 6). They add that a good understanding of the interplay of these capabilities is key to evaluate the impact of AI. Slightly different capabilities were identified by Brock and Wangenheim (2019). Instead of People and Processes, they propose Strategy and Security as key capabilities alongside Technology and Data (Brock & Wangenheim, 2019, p. 116).

Expanding upon these studies, further research was undertaken to explore and substantiate the needed capabilities. Focusing on organizational capabilities, Weber et al. (2023) identified four dimensions: AI project planning, co-development of AI systems, data management and AI model lifecycle management. Regarding AI project planning, they emphasized the need to identify and evaluate suitable use cases, as 76% of companies struggle to scale AI successfully across their organizations (Reilly et al., 2019, p. 3). In this context, understanding the specific characteristics and limitations of AI models was mentioned as an important prerequisite to better understand the possibilities of AI and to manage expectations (Weber et al., 2023, p. 1555). As pointed out by the researchers: "There is much value in professionalizing and democratizing the process of use case generation to collect fruitful AI use cases" (Weber et al., 2023, p. 1556). They state that each use case should have a clear value proposition to enable the company to prioritize and pursue the ones with the highest value creation potential.

Another important aspect that was frequently mentioned is the collaboration between technical and business functions (Uren & Edwards, 2023, p. 7). It is pointed out that data scientists need to work closely together with business experts to drive AI projects successfully (Akkiraju et al., 2018, p. 6). Including domain experts into the process plays an essential role as AI solutions can have a major impact on existing work structures. Co-development between IT professionals, data scientists and domain experts can help to reduce potential fear of AI systems and help to drive the implementation more effectively (Weber et al., 2023, p. 1556).

Mentioned in various scientific papers, data management and capabilities seem to be a key factor for successful AI adoption (Brock & Wangenheim, 2019; Uren & Edwards, 2023; Weber et al., 2023). The authors mention that many use cases do not make it into productive implementation due to insufficient data quality or unsuitable data structure (Weber et al., 2023, p. 1557). Implementing a clear governance process for the collection, storage and curation of data is therefore a vital step to ensure the successful application of these use cases in a real-world setting (Uren & Edwards, 2023, p. 6). In this context, it is important for organizations to not only assess their existing data management strategies but also to anticipate and plan for the evolving demands of future AI solutions (Weber et al., 2023, p. 1557).

Implementing new AI solutions raises questions about their long-term management in large corporations. Weber et al. (2023) highlight that this issue is particularly relevant given the iterative development process of AI models, involving multiple tests and adjustment cycles. Coordinating these releases is critical, especially when AI models are implemented across various organizational units, each potentially requiring minor or major adjustments. Consequently, organizations must ensure that AI solutions are integrated into their IT landscape and aligned with their business processes. The ability to manage these steps in the AI models' lifecycle is thus considered a crucial factor in successfully introducing AI into an organization (Weber et al., 2023, p. 1557).

Across many publications, special attention was given to the people dimension. Analyzing capabilities for digital transformation, Blanka et al. (2022) identified employee transformation competency to be crucial for driving digitalization in companies. They observed that employees as well as managers need to have a basic understanding of the potential of digitalization in order to effectively leverage it for organizational innovation (Blanka et al., 2022, p. 10). To develop these skills further and to be ready for the introduction of AI, broad training measures are required (Lee et al., 2019, p. 8). According to the authors, AI experts are rare. Offering training to the workforce could therefore help to improve understanding and support the digital transformation process. This also seems to be applicable for Generative AI (Feuerriegel et al., 2023, p. 6). Expanding the knowledge about the technology in the organization could

help identifying use cases and increase the overall readiness to integrate Generative AI into everyday business processes. To overcome organizational resistance and successfully scale AI applications, Lee et al. (2019) propose the following five step process model, as displayed in **Figure 6**.

Starting with executive pilot projects, the authors argue that companies should focus on smaller applications first. This supports the goal to acquire knowledge and to celebrate success stories early on. Based on these success stories, use cases can then be scaled to larger projects. This is accompanied by the formation of dedicated AI-teams which drive the topics forward. The authors acknowledge that this is likely not possible for smaller companies or startups which should consider support from 3^{rd} party providers. Besides the broad AI-training, the development of an AI-strategy is highlighted as a crucial aspect. While Lee et al. (2019, p. 8) focus on the importance of data in this context, it can be argued that a more holistic view is required as outlined by Brock and Wangenheim (2019) and Uren and Edwards (2023). Finally, Lee et al. (2019) highlight the importance of communication to all stakeholders. Due the novel character of AI and corresponding uncertainties, they point out that communication about the advantages and potential challenges is needed. They add that this could also help to find new, innovative ideas for AI when discussing use cases with customers (2019, p. 9).

4. Methodology

The research design of this thesis follows an exploratory case study approach as outlined by Yin (2018) with a clear methodological path, which is based on a comprehensive literature review as described in Chapter 3. For background information on the technology and related scientific research, two major streams have been followed. Firstly, the development of artificial intelligence, starting with the first conceptual ideas in the 1950s up to the latest advances in the field of Generative AI. Secondly, the review of scientific research on artificial intelligence and its value creation opportunities. Given the novelty of Generative AI, most available concepts and models have been written and published before its introduction, dealing with the available AI at its time. The following case study therefore offers a great opportunity to add new insights to the research field.

Supported by the literature review, an exploratory case study design is applied, an approach especially feasible for phenomenon which have not yet been comprehensively researched (Yin, 2018, p. 15). According to Yin, a case study is an empirical method to investigate a phenomenon or "case" from multiple angles to create an in-depth understanding. It is especially useful when the phenomenon can only be analyzed within its real-world context and its boundaries are not clearly visible. The goal is therefore to create an in-depth understanding of the subject and its current application which should allow for analytic generalization (Yin, 2018, p. 21). Doing so, the case study follows a single-case design which allows for a focused study of the selected subject and rich data collection (Yin, 2018, p. 49).

4.1. Case Study Design

In the search for a suitable object of investigation, two essential questions were answered. First, does the case provide a good basis for exploring the value creation potential of Generative AI? And secondly, is it possible for the author to gain valuable insights for the research within the scope of his possibilities? By defining the company Siemens as the subject of the study, both questions could be answered in the affirmative. On the one hand, the company with its multinational setup and different business areas offers ideal conditions to gather diverse perspectives on Generative AI. On the other hand, as an employee of the company, the author is able to gain insights that are not available to external parties. The latter circumstance in particular promises to provide valuable insights into the company and its approach to new technologies. Moreover, it offers a unique opportunity to validate scientific theories and concepts using a real-life example.

Founded in 1847, Siemens is a multinational corporation headquartered in Munich. Siemens sees itself as a technology company with the mission to "provide technologies that improve quality of life and create lasting value for society" (Siemens, 2024). The company focuses on four core businesses: Digital Industries, Smart Infrastructure, Mobility and Siemens Healthineers. Further businesses include Siemens Financial Services and its portfolio companies Innomotics and Siemens Logistics (Siemens, 2023, p. 85). Corporate functions include among others Strategy, Compliance, Supply Chain Management and Global Business Services as well as IT, Human Resources, Controlling and Communications. In fiscal year 2023, ending on September 30, 2023, Siemens generated revenue of € 77.8 billion from continuing operations with a corresponding net income of € 8.5 billion. As of September 30, 2023, around 320,000 people were employed by the company worldwide (Siemens, 2024).

The case study was able to collect information from almost all areas of the company. Only information from Siemens Healthineers had to be omitted, as it is managed as an independent company within the Siemens Group. In addition to interviewing managers and employees from Digital Industries, Smart Infrastructure and Mobility, participants were recruited from various business functions, including Human Resources, Global Business Services, IT, Strategy, Supply Chain Management, R&D as well as Technology.

4.2. Data Collection

Semi-structured interviews with key informants form the core of the study. These interviews provide helpful qualitative data and insights from various perspectives. The interview partners were selected specifically in order to obtain a broad view of the research subject. This "triangulation of subjects", as outlined by Rubin and Rubin (2012, p. 63), aims to bring together different ideas, perspectives, and opinions. In addition, when selecting the contact persons, attention

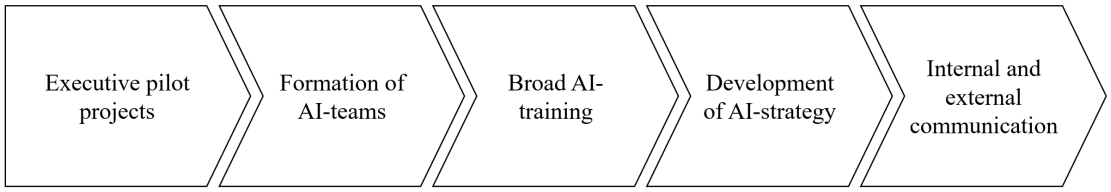


Figure 6: Five Step Process Model, based on Lee et al. (2019)

was paid to their background knowledge of the technology in order to generate high-quality input for the study (Rubin & Rubin, 2012). The interview partners were from various areas, roles and responsibilities, including CFOs, CIOs, Program Managers for Generative AI, Product Owners, Domain and Technology Experts as well as IT Professionals and Innovation Managers (see **Table 1**).

Interviews lasted between 24 minutes and 1 hour and 15 minutes. Due to separate locations and time zones of the audience, all interviews were held virtually via Microsoft Teams, using the latest technology for recording and transcription. To allow for an open exchange and unfiltered perspectives, the confidentiality of the meetings was agreed upon and all insights will only be used in an anonymous way.

A total of 23 interviews were conducted in the period from December 2023 to March 2024 with an average length of 43 minutes. MAXQDA, an external software designed for computer-assisted qualitative data analysis, was used to analyze the interviews. During the process, the transcripts were anonymized, cleaned, and organized. It was found that many transcripts contained erroneous sections and thus required manual correction of words. This was due to the fact that the software used for transcription, Microsoft Teams, did not correctly recognize the interviews that were conducted in German. Furthermore, due to the specific topic and many abbreviations used within Siemens, many technical terms were not correctly recognized by the software. Manual coding was therefore conducted with the help of the transcripts and video recordings. New interviews were arranged until it was ensured that all aspects were sufficiently covered, and no blind spots remained. Reaching this 'saturation point' which was first described by Glaser and Strauss (1967, p. 61) contributes to the depth and completeness of the research findings.

A guideline was drawn up for the interviews in order to give the discussions a clear structure and to facilitate the subsequent evaluation. **Table 2** shows the general structure and topics of the interview guide. In addition to the guiding questions, the interview partners were given space to address additional aspects and explain them in detail. This ensured that findings and topics not yet covered by questions were also recorded (Myers, 2020, p. 149). The interview guide was successively revised after each interview to take new aspects of the data collection into account. This helped to verify aspects raised by interview partners in subsequent interviews.

Besides the semi-structured interviews, other sources are used to supplement and enrich the case study. From the six sources of evidence, as outlined by Yin (2018, p. 113), the following are included in the thesis

- Interviews

  Interviews, as previously discussed, are conducted to gather valuable insights, perspectives, and explanations from individuals who possess key information. This process follows a structured interview guide containing predefined topics and questions. However, it also allows for flexibility to explore alternative inquiries and different areas of focus (Rubin & Rubin, 2012, p. 31).

- Documentation

  The data collection process for the case study involves systematically searching for relevant literature. This serves as a crucial component for future reviews and data triangulation. It comprises several types of data, including press releases, internal presentations, meeting minutes, reports, news articles, wikis, and information stored on SharePoint.

- Direct Observations

  Direct observations provide valuable additions to the mentioned interviews. These observations encompass various activities within the context of the research phenomenon, such as conferences, meetings, recorded sessions, training courses, and other relevant events. Furthermore, internal communities dedicated to the research topic offer insights into the phenomenon and serve as an additional source of information.

- Participant Observation

  The author of this thesis does not merely act as a passive observer but actively engages in exploring Generative AI within the case study's context. In his current role within the company, he delivers presentations on the topic, participates in conferences and meetings related to Generative AI, and engages in discussions with colleagues regarding the technology's ongoing development. As described by Yin (2018, p. 124), this active involvement presents both unique opportunities and risks, which have been carefully evaluated for their potential impact on the case study. Firstly, as an active participant, the author can gain practical knowledge in addition to theoretical insights from his firsthand experiences. Secondly, it enables the author to partici-

Table 1: Overview of Interviews

No	Role	Organization Type	Length	Date
1	Program Manager	Human Resources	55 mins	Dec 23
2	Program Manager	Mobility	52 mins	Dec 23
3	Product Owner	IT	44 mins	Dec 23
4	Domain Expert	Supply Chain Management	55 mins	Dec 23
5	IT Professional	IT	51 mins	Jan 24
6	Product Owner	R&D	54 mins	Jan 24
7	Domain Expert	Supply Chain Management	54 mins	Jan 24
8	CFO	Mobility	45 mins	Jan 24
9	Technology Expert	IT	27 mins	Jan 24
10	Program Manager	Global Business Services	42 mins	Jan 24
11	Technology Expert	IT	29 mins	Jan 24
12	Domain Expert	Mobility	1h 15 mins	Jan 24
13	Domain Expert	Mobility	52 mins	Jan 24
14	Program Manager	Technology	53 mins	Jan 24
15	CIO	IT	24 mins	Jan 24
16	Head of Technology	Digital Industries	33 mins	Jan 24
17	Domain Expert	Digital Industries	30 mins	Feb 24
18	Technology Expert	Technology	35 mins	Feb 24
19	Domain Expert	Smart Infrastructure	26 mins	Feb 24
20	Strategy Manager	Strategy	30 mins	Feb 24
21	Program Manager	Smart Infrastructure	40 mins	Feb 24
22	Program Manager	Smart Infrastructure	54 mins	Feb 24
23	Innovation Manager	Digital Industries	32 mins	Mar 24

pate in events and establish connections with individuals who might not have been included in the study otherwise. However, active involvement also poses significant challenges that require close monitoring (Yin, 2018, p. 124). Particularly, unintentional bias in supporting the research subject could lead to a subjective and skewed perspective. The author acknowledges this trade-off and takes utmost care to ensure an objective analysis.

The underlying goal of any research is to thrive for the highest possible quality. Its design should follow a logical sequence which allows other researches to analyze and verify its findings (Yin, 2018, p. 42). During the case study, the design quality was therefore measured and verified using three common tests. These are construct validity, external validity and reliability (Yin, 2018, p. 42). According to Yin (2018), construct validity aims at finding the correct measures to assess the underlying theoretical construct (Yin, 2018, p. 43). Using multiple sources, establishing a chain of evidence and reviews of the case study helped to increase the construct validity. The second test examines the degree to which the results of the study can be generalized to a different or broader context (Yin, 2018, p. 45). Grounding the research questions in appropriate theory, supports the external validity of the study. Ensuring reliability of the findings and conclusions, represent the final test (Yin, 2018, p. 46). To fulfil these requirements, the case study follows a clear approach and provides comprehensive documentation, allowing other researchers to potentially replicate the results.

4.3. Analytical Approach

As outlined above, the analysis of the data follows a six step approach of Braun and Clarke (2006, p. 15). This allows for a deep understanding of the gathered information and underlying patterns and provides a clear structure during the analysis. The first step involves the familiarization with the data. Structuring the entire data set, reading through the interviews and extracting first ideas and patterns, helps to lay the foundation for the in-depth analysis (2006, p. 16). In a second step, initial codes are generated. These codes summarize the meaning of a section, phrase or word and represent the building blocks for the analysis. They help to summarize

Table 2: Interview Guide

Introduction
Brief introduction and purpose of the study
Clarification of confidentiality and anonymity
Part 1: Background Information
Walkthrough of professional background and experience with AI
Current involvement in Generative AI
Part 2: Strategic Understanding of Generative AI
Comparison of Generative AI with previous AI models
Functionalities and limitations of Generative AI and its relevance for Siemens
Part 3: Value Creation Opportunities
Use Cases - insights, development and value add for the company
Application areas with the highest potential value creation
Combination of Generative AI with other technologies
Outlook on future value creation opportunities
Part 4: Expanding AI within Siemens
Requirements for successful implementation and scaling across Siemens
Potential challenges and barriers (technical, organizational, etc.)
Part 5: Ethical and Security Considerations
Ethical considerations when introducing Generative AI
Cyber and Data Security
Conclusion & Wrap-Up

and classify the data and provide more transparency over the findings (2006, p. 18). The third step deals with the identification of themes. Based on the identified codes, patterns are extracted and merged to larger groups with the same meaning. These themes help to better understand ideas across the data set. Starting with the analysis, an inductive approach was followed allowing themes to emerge from the data itself (Myers, 2020, p. 210). In the fourth step, the identified themes are reviewed to ensure consistency and quality (Braun & Clarke, 2006, p. 20). This step included going back to the initial codes and verifying their validity and fit to the selected category as well as revealing potential biases and misunderstandings. In addition, the analysis as a whole was analyzed for ambiguities and overlaps. This proved to be a valuable step, as familiar patterns from observations in the company and aspects from the literature emerged. Following a deductive approach, these were considered as follows: In places where the data showed strong similarities with existing concepts, the terms were taken from the literature. For new topics, own terms were chosen. This approach made it possible to uncover previously unknown themes in relation to Generative AI and, at the same time, made it easier to compare recurring themes with existing models and concepts (Myers, 2020, p. 210). Step five of Braun and Clarke's (2006) approach is closely linked to the previous one and deals with the definition and naming of themes. In this step, the emerging themes were once again refined to ensure that the underlying data was represented correctly (2006, p. 22). In this step, subcategories were identified for several themes which allowed for more transparency and a better understanding of the identified patterns. **Appendix 1** provides an overview of the themes and their occurrence in the interviews. The sixth and final step involves the presentation of the results. This will be done in Chapter 5, followed by a discussion in Chapter 6 which aim at providing a comprehensive insight into the findings and critical analysis of the data (2006, p. 23).

In total, 1,139 segments were identified in the raw data which were assigned to 429 codes. These were divided categories and subcategories, which were subsequently categorized into five main themes. **Appendix 2** shows the process from raw data, through open codes and categories to the main themes for a selected sample.

5. Results

The information for this case study was collected over 12 months from March 2023 to March 2024. During this period, the author participated in various meetings in which Generative AI aspects were presented and discussed. In addition, the author became a member of various groups actively involved in the introduction and dissemination of Generative AI in the company, regularly exchanged ideas with colleagues on new findings and participated in internal and external conferences. Furthermore, the author himself gave presentations and organized workshops to create awareness for Generative AI and to gain further insights. This was done regionally, in Asia Pacific, but also globally in management meetings with representatives from different business areas and functions. Moreover, in order to better understand the technological aspects of Generative AI, the author initiated a pilot project together with a team of data scientists (**Appendix 3**). The core of the case study form interviews with experts from various business areas mentioned earlier. Finally, the research work is rounded off by the collection of internally and externally available documents.

5.1. Comparative Analysis: Generative AI vs. Previous Models

A key question that arises when it comes to Generative AI is what makes the technology so different and interesting for companies. Two key factors can be identified from the analysis: the technological innovation itself and its high usability compared to previous AI models. In terms of technology, three factors emerged from the data: improved textual understanding, learning speed and flexibility of the models. Interview partners pointed out that this makes it possible to realize use cases that were too complex with previous models

due to the effort required for adaptation and training. Illustrating this point, interview partner 11 (IP-11) stated:

> *"Foundation models allow us to solve many more use cases than before because we no longer have to collect data and train any model and that is [...] very disruptive for the AI world because that was always the number one problem before - where do I get my data?" (IP-11, Technology Expert)*

The data also provides a clear picture with regard to usability. Interview partners state that interacting with artificial intelligence is now remarkably simple. The conversational user interface, which allows anyone to execute commands via prompts, is named as a key driver in this context. This empowers a wide range of employees to try out the tool without having to worry about complex control parameters or the underlying technology. As IP-7 puts it:

> *"The biggest added value of the interface itself is that you really only need text and speech to interact with [the] tool" (IP-7, Domain Expert)*

The data suggests that this will open up new application possibilities in the future, as the models can now be used by employees without any technical background (*IP-16, Head of Technology*). In the future, for example, users could interact with a database for which SQL knowledge is still required today (*IP-9, Technology Expert*).

5.2. Value Creation Clusters

The collection and analysis of Generative AI use cases was an essential component of the case study. In addition to the pilots mentioned in the interviews, other internal sources of information were used to obtain a complete picture. This resulted in three perspectives from which potential value creation can be viewed. The first perspective involves categorizing the use cases according to the way in which the technology is used. The second perspective focuses on value creation drivers, i.e., use cases offering new value creation opportunities or improving existing structures. The third perspective divides the value creation opportunities in organizational units, allowing companies to identify key areas within their structures more easily. **Figure 7** illustrates the individual categories and subcategories.

Looking at value creation clusters, from the data collected, it became apparent that most of the use cases can be assigned to the **assistants'** area. A further subdivision was therefore possible and allows for a more precise differentiation of use cases. It must be mentioned that during the case study, various approaches were identified at Siemens for the clustering of use cases. These were used as the basis for the initial categorization, revised and refined over the course of time. This resulted in a clear picture of the assignment of the individual use cases.

Within the first cluster, the area **data synthesis** deals with use cases that focus on extracting, analyzing, and gaining knowledge from data. Generative AI is used to gain new insights that could not previously be achieved in this form or only with considerable additional effort. Examples include a Generative AI-supported tender analysis at Siemens Mobility to classify requirements and identify risks, the identification of common parts across multiple IT systems in Siemens Procurement or the analysis of employee comments as described in detail below.

Example: Employee Surveys

> Siemens regularly asks employees about their satisfaction as part of a global campaign. This survey also includes free-text fields in which employees can write comments. With the help of Generative AI, a sentiment analysis can be carried out on all comments to determine whether the message was written in a positive, neutral, or negative tone. In this way, critical topics such as the risk of burnout can be identified without being explicitly mentioned in the text (IP-1, Program Manager). Consequently, Generative AI proves beneficial by allowing for an analysis without the need to set up and train a dedicated AI system specifically for sentiment detection.

Another potential use case that was regularly mentioned was the adaptation and training of foundation models to include Siemens specific content. While no advanced use cases were identified in the interviews, multiple research projects were mentioned, and many interviewees highlighted the potential for future applications. As expressed by one participant, the lower training effort compared to previous approaches promises a major advantage, as the models would already have a sufficient understanding of human language (IP-19, Domain Expert).

The second cluster, Generative AI-based **assistants**, contained by far the most application examples. The aim of these systems is to prepare and process existing information, making it readily available to the user. The subcategories are based on the complexity of the respective use case; factors include the type of data processed (static or dynamic), the number of systems connected, the query type (one-shot or staged), and the level of required human interaction. The necessary user knowledge also plays a role. While the simplest use cases require no prior knowledge, other systems are designed for IT professionals and data scientists.

With **information retrieval**, users are generally interested in finding information more quickly. These "factual chatbots" can access a variety of data sources, ranging from internal company wikis, guidelines, and circulars to department-specific knowledge databases. From a technological point of view, these systems use Retrieval Augmented Generation (RAG) in combination with a licensed large language model, approved for confidential internal information.

Use cases in the second subcategory **interplay with systems** require a close link between the LLM and the data source. The most common example cited was using Generative AI as a user interface to query internal databases. Here,

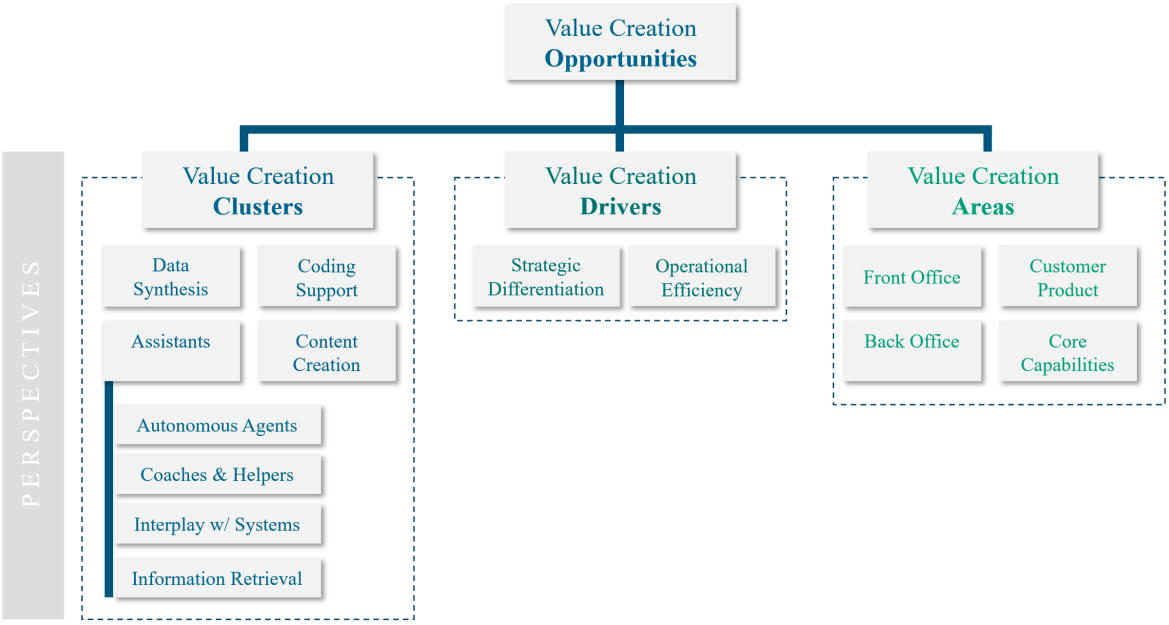


Figure 7: Value Creation Opportunities

Generative AI converts user commands in natural language into SQL commands to retrieve information. One interviewee emphasized that this could enable faster, more individualized queries in the future, decreasing the need to create new dashboards for end users (IP-4, Domain Expert). Other examples include combining Generative AI with Robotic Process Automation (IP-3, Product Owner), using Generative AI alongside SAP inputs (IP-7, Domain Expert), or evaluating application documents in the Siemens-wide job portal using Generative AI (IP-1, Program Manager).

The data revealed that the main focus of assistant systems is on use cases that actively support users in their tasks. These **coaches & helpers** accompany the user along the value chain in various process steps and influence the desired results. Added value is expected in two areas in particular. In the use of external tools such as Microsoft Copilot and in the development of internal, domain-specific assistants, such as a chatbot for service technicians in the area of Smart Infrastructure (IP-21, Program Manager). Further application examples include the development of chatbots to prepare and practice negotiations in purchasing (IP-7, Domain Expert), the use of Generative AI in interview processes (IP-9, Technology Expert) or support for process reviews as part of internal audits (IP-12, Domain Expert).

While the majority of the identified use cases envisage the above-mentioned assistance role for Generative AI, individual interview partners see the possibility of using the technology to **automate** entire processes. This usually includes the interaction of several software applications and systems, in the context of which Generative AI represents the interface to the user. Illustrating this point, an interviewee mentioned a Generative AI agent that makes it possible to classify customer inquiries automatically, extract the necessary information, update the linked systems, and write a response to the customer (IP-10, Program Manager). It was pointed out that while this was already possible in theory, in reality the decisions would still be made by humans due to the maturity of the use case, the specific features of the technology and the associated risks.

Another large cluster comprises use cases that can be summarized as **coding support.** As a technology company, Siemens employs software in a wide variety of forms and areas. The ability of Generative AI to generate code in different programming languages was therefore seen as a great lever to increase productivity and improve existing software solutions. One example mentioned in the interviews and communicated externally by Siemens as a lighthouse project revolves around the Siemens Industrial Copilot, as described in detail below.

Example: Siemens Industrial Copilot

> Siemens offers a variety of solutions for factory automation, including control systems for machines and processes. In this use case, Generative AI is deployed to write PLC (Programmable Logic Controller) code for these machines. A distinctive feature mentioned is the ability to translate and improve code from other programming languages. Additionally, the software leverages Generative AI to identify bugs and suggest solutions, significantly reducing programming effort and potential downtime in factories.

Another example of Generative AI used within Siemens' software solutions is its implementation in the Mendix low-code platform. In collaboration with external partner AWS, Siemens extended the program to further support users in the

development, validation, and optimization of Mendix applications. In addition to these two customer-facing solutions, many interviewees reported added value in developing internal applications. One example mentioned by a respondent was Generative AI assisting in the development of an automation script for a building's energy supply (IP-19, Domain Expert). A common theme from the interviews was that the coding environment's focus was less on individual use cases and more on transforming the way programming is done. Many respondents see great added value in the ability to generate and improve code, as well as in using assistants like GitHub Copilot (IP-6, Product Owner).

The last cluster can be described as content creation, where the focus is on generating new insights and data with the help of Generative AI. Common examples include its use in marketing for writing news articles or in human resources for generating job advertisements. In addition to these general applications, specific Siemens use cases were identified. For example, one interview partner reported the use of Generative AI as part of the innovation process to find new alternative materials for products (IP-7, Domain Expert). Another participant reported using Generative AI to translate 3D models of metal parts into code for milling machines (IP-18, Technology Expert).

In summary, the collected use cases indicate a broad range of applications for Generative AI. In addition to simple models for information retrieval, the technology also offers added value for complex internal processes as well as customer-facing applications. In this context, the identified value creation clusters provide a valuable starting point for better understanding the diverse potential of these use cases.

5.3. Value Creation Drivers

The data analysis revealed two main drivers of Generative AI use cases, as highlighted in **Figure 8**.

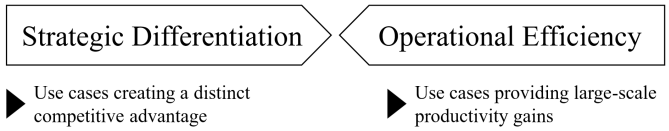


Figure 8: Value Creation Drivers

The first group can be broadly summarized as **strategic differentiation**. These use cases focus on gaining a unique competitive advantage by leveraging resources exclusive to the company, enabling differentiation from the competition. These resources include internal company data (such as customer information), process knowledge, and the IT infrastructure itself. This group also encompasses the development of new products or the expansion and application of Generative AI capabilities within existing products. Examples include the Industrial Copilot, which Siemens uses to strengthen its own software applications, and the integration of Generative AI into the Mendix low-code platform.

Internally, this can involve developing technology stacks and platforms to accelerate the rollout and distribution of specific use cases. For instance, the Global Shared Services division developed a Generative AI platform that enables departments worldwide to adopt simple use cases without needing to invest in the technology themselves (IP-10, Program Manager). This creates synergy effects, as the internal solution requires development only once and can subsequently be rolled out across the entire organization.

The second driver is productivity improvement. The data and interviews revealed that the majority of current use cases fall into this category, less focused on unique solutions and more on leveraging the technology effectively. This includes making solutions broadly accessible to as many users as possible to improve **operational efficiency** on a large scale. Participants frequently mentioned faster information retrieval as one of the biggest levers. Others listed quality improvements, faster execution of processes, and risk minimization as productivity-boosting factors.

5.4. Value Creation Areas

One finding that emerged during the case study is the company's need to cluster potential use cases not only according to its technological background, but in particular according to company-specific segmentation. This was particularly noticeable when analyzing company presentations and participating in meetings on the possible areas of application of Generative AI. Depending on the focus, different perspectives can be favored in the various company divisions. **Table 3** shows three distinct views on use cases which are used within the company. The first view is built around functional areas within the company and allows interested parties to differentiate Generative AI use cases according to an established company-wide used segmentation. The second view is built around business capabilities which describe what businesses do to achieve a specific purpose. Driven by processes and not by organizational structures or technology, this view is often used in IT for strategic analysis. Lastly, clustering in opportunity areas provides a more generic view which is in particular suitable for gaining a faster understanding of Generative AI application areas. Originally proposed by Gartner (2023), Siemens adopted the model for its own organization. It distinguishes between front and back-office tasks, product, and services as well as core capabilities of a company.

5.5. Exploration and Scaling

While focusing on the value creation opportunities of Generative AI, the case study identified several supplementary elements. Many interviews provided insights into the current state of Generative AI at Siemens and factors that should be considered for a successful company-wide rollout. Analyzing the data, two major patterns emerged:

Firstly, Siemens' current state regarding Generative AI can be described as an **exploration phase**. This is driven by the technology's novelty and Siemens' approach to the topic over the last year. **Figure 9** shows the key aspects of this exploration phase that were mentioned during the interviews and observed during the case study.

Table 3: Value Creation Areas

Functions	Business Capabilities	Opportunity Areas
Legal	Strategic & Information Management	Product / Services
Customer Service	Sales & Marketing	Front Offce
Operations	Innovation & Lifecycle Management	Back office
Finance & Strategy	Supply Chain Management	Core Capabilities
Product and R&D	Manufacturing and Production	
Sales & Marketing	Service Management	
Quality Management	Project Management	
Human Resources	Enterprise Services	
Supply Chain Management		

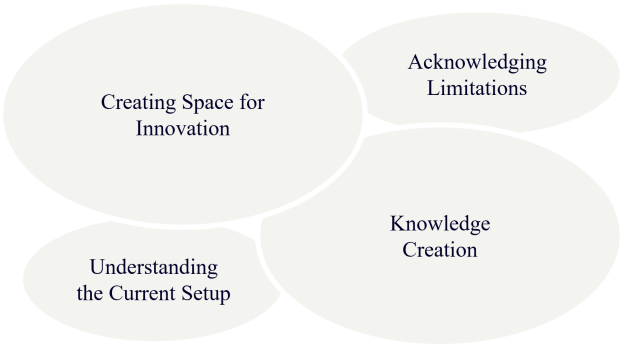


Figure 9: Key aspects of the exploration phase

A consistent pattern that emerged across interviews was **creating space for innovation.** Interview partners frequently noted that they were given the resources and time necessary to explore this new technology in depth. While this seems logical for technology departments previously involved in NLP development, many business areas were also granted the resources to participate in the early ideation stage. Remarkably, many participants, particularly domain experts, highlighted that their exploration of Generative AI was driven by their own intrinsic interest. Based on this personal motivation, they were given the freedom to explore the topic within a company context (IP-13, Domain Expert).

The second theme contributing to the exploration phase is knowledge creation. Findings suggest that many use cases are far from deployment or are already obsolete due to the field's rapid technological advances over the last year. However, as emphasized by many interview partners, the knowledge gained during their development is a valuable asset for the future. Additionally, several participants outlined the need to invest in the area of Generative AI to stay competitive. As IP-15 puts it:

> *[. . .] we need to start building up resources as quickly as possible, because once the topic takes off, even with competitors, it will be all the more difficult to recruit new people (IP-15, CIO).*

Lastly, two more themes emerged in several interviews. **Understanding the current setup** summarizes feedback that companies should carefully analyze their current pain points to apply new technologies such as Generative AI effectively. Participants pointed out that this analysis should consider both the quality of existing data and potential bottlenecks (IP-5, IT Professional), as well as the business processes where improvement offers the greatest added value (IP-10, Program Manager). Finally, **acknowledging limitations** highlights the fact that the introduction and exploration of new technologies comes with various uncertainties. One uncertainty is the rapid development of Generative AI itself, making it difficult to commit and invest in a specific solution, as it could become obsolete in a few months. Technology-inherent limitations such as hallucinations, the limited explainability of its outcomes as well as regulatory and legal considerations were named as additional constraining factors (IP-16, Head of Technology).

Besides the outlined area of exploration, interview partners expressed various requirements and challenges which can be summarized under the term **scaling**. It became apparent that for large corporations, the path from individual pilots to widely accepted approaches and processes is a critical hurdle that must be overcome to achieve sustainable added value from new technologies like Generative AI. To facilitate a better understanding of the various aspects, the results are categorized into the following dimensions:

- People
- Processes
- Technology
- Data
- Organization
- Strategy
- Communication
- Timing

While slight overlaps are possible between these dimensions (such as People and Organization or Timing and Strategy), there are distinct differences. This is also evidenced by the fact that over 500 code segments from the interviews could be assigned to the area of scaling. The main findings from the analysis are therefore presented below.

The **people** dimension deals with the influence of Generative AI on employees within the company. A dominant theme that emerged repeatedly was the belief that Generative AI will have a major impact, requiring reskilling and upskilling of the workforce (IP-1, Program Manager). Participants noted challenges stemming from the technology's rapid pace of evolution. While some colleagues adapt quickly, interviewees expressed concern that others, often from the older generations, might struggle to keep up. This could foster a negative attitude towards Generative AI, further fueled by fears of job displacement (IP-2, Program Manager).

Another hurdle identified was the lack of experts with both business knowledge and deep understanding of the technology, making value-adding applications difficult to identify (IP-3, Product Owner). Additionally, due to widespread media coverage and heightened expectations, maintaining employee motivation as initial ideas turn into concrete use cases becomes challenging (IP-5, IT Professional).

Several key factors for overcoming these hurdles were identified. Primarily, awareness must be fostered through training, workshops, or coaching, accompanied by clear expectation management. One interview partner emphasized that the psychological impact, and thus the human factor, cannot be underestimated (IP-8, CFO). Cultural differences must be considered to prevent negative consequences from the introduction of Generative AI. A simple but illustrative example is the AI-supported recording and transcription function in Microsoft Teams, which summarizes meeting content, analyzes it, and defines action items. This could lead to employees feeling uncomfortable and contributing less to future meetings.

Finally, talent management is crucial for successful scaling within the organization. Identifying key internal resources is important to recognize gaps and, if necessary, fill them with external candidates. In this context, increased cooperation with universities was mentioned as a way to attract young talent to the company (IP-15, CIO).

In terms of **processes**, the data analysis revealed that the introduction of Generative AI creates challenges similar to those faced with other technologies. This can be attributed to the complex and heterogeneous process landscape within companies like Siemens. The diverse business areas, specific requirements, and sometimes historically grown processes limit the company-wide scaling of new technologies. Interviews indicated that participants view more uniform processes and standardized approaches as major levers for scaling (IP-14, Program Manager). In this context, particular importance was placed to new software solutions whose setup would offer the possibility to consider the requirements of Generative AI at an early stage (IP-5, IT Professional).

In terms of the **technology** itself, several interesting insights emerged. It became clear that many use cases failed to meet initial expectations, primarily due to the technology's current maturity and limitations for complex business processes (IP-1, Program Manager). A key aspect is the previously mentioned issue of hallucinations. Unlike previous models that provided deterministic answers, Generative AI responses are probabilistic (Radford et al., 2019, p. 2; OpenAI, 2023). It generates outputs by calculating probabilities based on its training data. The technology's maturity also presents a significant organizational hurdle. Companies must identify and evaluate different providers and solutions before approving them for internal use. This process requires time and resources and can become a bottleneck due to the rapid pace of the technological development. The data revealed several elements to address these challenges: First, Companies should strive to remain vendor and model independent. Open AI, Microsoft, Google, AWS, and open-source providers constantly release newer and better Generative AI solutions. As mentioned by IP-6, internally developed platforms should be as independent as possible to benefit from these enhancements. Second, technological stacks should be shared within the company to avoid duplicate approaches and reduce initial development efforts (IP-3, Product Owner). Third, companies must carefully weigh whether internal development is necessary or if solutions can be purchased externally. This includes cases like in-house training of large language models or use cases that might be covered by future Microsoft Copilot functionalities (IP 16, Head of Technology).

The **data** dimension plays a significant role in the successful implementation and scaling of Generative AI. This stems from the expectation that companies can achieve a competitive advantage through leveraging their own data. However, many participants expressed the need to structure and connect data within the company for meaningful use (IP-5, IT Professional). Additionally, insufficient data quality in certain areas poses a challenge. The findings suggest that a clear data strategy addressing the aforementioned aspects is necessary for successful scaling of the technology. As one interview partner mentioned, it could be beneficial to appoint a Chief Data Manager to emphasize the importance of the topic and guide it in the right direction (IP-8, CFO). The Siemens Data Cloud, an existing approach to data sharing and streamlining data processing, was cited as a positive example (IP-1, Program Manager). Overall, all participants emphasized the importance of this dimension.

Considering the complexity of large companies, the **organizational impact** of groundbreaking technologies like Generative AI can create significant challenges. This is the case for Siemens, as evidenced by the data analysis. A recurring challenge frequently addressed by the interview partners is the potential for a silo mentality, where multiple teams might work on related topics independently (IP-22, Program Manager). While decentralized structures offer individual business units freedom, they can hinder the efficient introduction of new technologies. Therefore, clear structures, a common strategy, and the sharing of technological approaches

are considered essential for successfully scaling Generative AI use cases (IP-4, Procurement).

During the exploration phase mentioned at the beginning, the aim is to build up knowledge and try out all facets of the new technology. However, converting these findings into added value for the company requires a clear **strategy**. Findings suggest this is a challenge regarding Generative AI, as there is a lack of clear metrics to measure the success of individual use cases. Especially regarding productivity-enhancing use cases, evaluation in absolute figures becomes difficult as pointed out by one interview partner (IP-10, Program Manager). Findings suggest that when developing Generative AI solutions, prioritizing business areas where the technology has the greatest potential impact seems beneficial. This aligns with recommendations from interview partners. For example, the engineering area at Siemens Mobility was identified as a potential focus due to its large share in project execution and lack of skilled railway engineers. Productivity improvements in this area could therefore yield substantial improvement potential (IP-2, Program Manager). This aligns with a key finding when analyzing individual use cases. Generative AI offers a myriad of possible applications in a company like Siemens. Consequently, it was emphasized that a considered approach to resource allocation, including deliberate decisions on individual use cases, is essential for optimal value (IP-8, CFO).

The vast attention surrounding Generative AI suggests the need for clear internal **communication** within the company. Interview partners pointed out that widespread awareness is key for effective cross-departmental use of the technology. An interesting pattern emerged during the case study: In the initial months, the desire for exchange and communication solidified into cross-departmental communities. These communities used internal platforms like Microsoft Viva Engage (**Appendix 4**) and Teams (**Appendix 5**), hosting regular information events. Existing AI structures were expanded, and their communication efforts were strengthened to meet the immense thirst for knowledge among employees and managers. One example is the "AI Attack" community, an IT group focused on streamlining the development and operation of AI solutions. After ChatGPT's release, a separate, secure internal Generative AI Platform was established in partnership with Microsoft (see **Figure 10**).

This platform enables employees to use the technology without fear of exposing confidential data and was accompanied by an information campaign, significantly contributing to the company-wide distribution. Other communities formed in areas such as human resources, further contributing to knowledge transfer and idea exchange.

In recent months, a new trend has been observed that fits in with the previous findings, namely the establishment of cross-sector initiatives. These reflect the need and desire to bundle use cases, create more transparency and drive the topic forward strategically. Due to the size of the company, these were not only set up at corporate level but also in the individual business units in order to take account of individual needs and priorities. One pattern, expressed by almost all interview participants, was the desire to create transparency around individual use cases. This reflects a pre-existing need within the company, as efforts to gather and showcase AI use cases were already underway. The emergence of Generative AI further amplified this need, leading to the creation of new use case collections across different areas of the company. One example at the corporate level is the so-called "Innovation Radar", as shown in **Figure 11**. The radar shows use cases from all areas of the company and stages of development, ranging from initial ideas to lighthouse projects. The interactive display allows the user to filter by different views, such as functions, objectives, or application clusters. As part of the case study, this overview was used alongside other area-specific use case collections to verify interview results and gain a better understanding of use case examples.

Finally, a theme that can also be seen as a subtopic of the other dimensions is **timing**. This dimension encompasses all observations regarding the appropriate sequencing and approach for the introduction of Generative AI. Interview data revealed diverse perspectives on this point, which can be divided into two areas. While the company often employs a well-defined structure and milestones for piloting and introducing individual projects, such as the Siemens Industrial Copilot, it faces challenges in implementing the technology at scale. As mentioned previously, the complex systems and processes were frequently cited as reasons for this. With regard to internal productivity increases, interviewees pointed out that a few impactful use cases could generate widespread employee enthusiasm for the technology. This aligns with the author's observations: while some employees are proactively integrating Generative AI into their daily work, the majority are not yet actively using it (**Appendix 6**).

5.6. Additional Observations

Further insights can be extracted from the conversations and data collected. A major topic area revolves around the risks associated with Generative AI and possible mitigation strategies. The Siemens legal and compliance department issued guidelines at an early stage regarding the use of external services such as ChatGPT. These included rules relating to sensitive information, illegal or unethical use, the protection of property rights, data privacy, and export control. Potential discrimination, biases, reputational risks, and questions about accountability for decisions made by Generative AI. Additionally, some participants expressed concerns about potential negative effects from anticipated increases in productivity. While many point to the skilled worker shortage and the need to streamline routine processes, others foresee more far-reaching consequences. As IP-1 puts it:

> *"I like the efficiency gain of it, but I also dislike the efficiency gain of it (IP-1, Program Manager)"*

Various mitigation measures were mentioned to minimize risks. The most frequent was "keeping a human in the loop", meaning that critical decisions should continue to be made or overseen by humans. Using Generative AI as a co-pilot

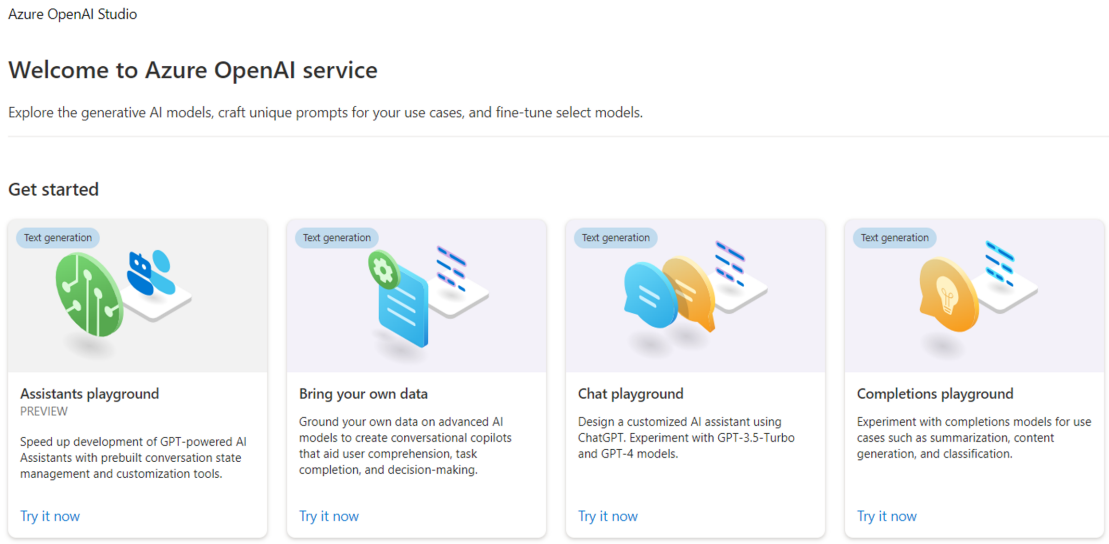


Figure 10: Siemens' internal Generative AI platform

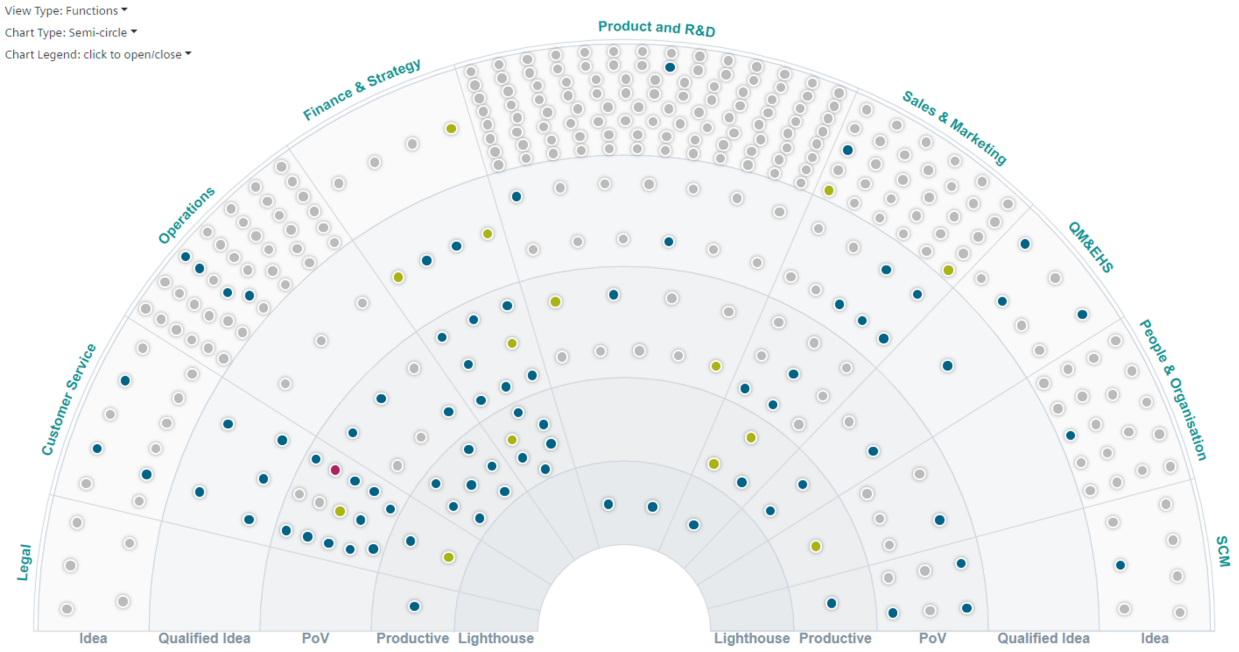


Figure 11: Siemens' Innovation Radar for Generative AI

(though this term has been adopted by Microsoft and others for their own solutions) was frequently mentioned in this context (IP-16, Head of Technology). From the interviewees' perspective, the use of synthetic data can reduce the risk of copyright lawsuits. And while the EU Artificial Intelligence Act is seen as a further challenge, it also provides guidance for future developments in the area of Generative AI.

Finally, two additional aspects from the interviews should be mentioned. Firstly, interviewees agree that Siemens can be seen as a pioneer in implementing Generative AI within its industry. This is attributed to the early provision of infrastructure (such as the internal Generative AI platform), which allowed employees from all business units to develop use cases. The associated knowledge building is considered valuable in initiating discussions and developing the framework for further implementations. Additionally, top management quickly embraced the topic, leading to regular discussions at internal events. External presentations, such as the Industrial Copilot at the Hannover Messe or Roland Busch's CES keynote speech, underline the importance of the technology from a corporate perspective. Lastly, another insight emerged: many interview partners see a kind of hype around Generative AI. This is particularly due to the low barrier to entry, though implementing value-adding applications for the company remains challenging. As IP-5 puts it:

"I think we all still have some way to go. On the one hand, learning how it works, but also a certain tolerance for frustration. [. . .] that will definitely come when the initial hype has died down and you realize that it's a bit more complicated to generate all this stuff. (IP-5, IT Professional)

6. Discussion

This case study explores the value creation potential of Generative AI within a large, multinational corporation. Spanning a one-year observation period and drawing on 23 in-depth interviews with experts and managers across various business areas, it offers a comprehensive view of the processes and challenges associated with introducing this transformative technology. The following section will delve into the key findings and their implications for both researchers and practitioners. This analysis will address the central research questions and highlight the study's theoretical and practical contributions. The chapter will conclude with a critical reflection, acknowledging the study's limitations and outlining potential avenues for future research.

6.1. Technological Innovation and Usability

When OpenAI launched ChatGPT in November 2022, the public was able to experience the advances in the field of artificial intelligence for the first time. Based on the outlined work by Brown et al. (2020), Radford and Narasimhan (2018), and Vaswani et al. (2017) and many others, ChatGPT showed what was possible with the help of "artificial intelligence". Some studies pointed to the possible significant influence of LLMs on different occupational groups (for example, Eloundou et al., 2023, p. 11) while others recognized the ability of Generative AI to take on creative tasks that previously could only be performed by humans (Feuerriegel et al., 2023, p. 1). Exploring **Research Question 1** on how Generative AI differs from previous models in terms of architecture and functionality, two key findings emerge from a business perspective. Firstly, the technological innovation enables companies like Siemens to explore new application areas and thus opportunities to generate additional business value. This is particularly important for existing products and software solutions that can be improved with the help of Generative AI. Examples that have been outlined in the case study are the Siemens Industrial Copilot integrated in Siemens' automation environment or Siemens' low-coding platform Mendix that incorporates Generative AI capabilities. While these are customer-facing products, the adaptability of Generative AI appears to be useful for internal applications as well. Use cases can be realized with significantly less effort, i.e., with a lower investment volume and in less time compared to previous models, thus enabling more business areas to deploy AI solutions to streamline their processes. Although long-term productivity improvements are not yet demonstrable, the case study has shown that efficiency gains are anticipated across all business. This starts with information retrieval systems which reduce the time spent for searching specific information and continues with assistant systems which actively support the user and help improve quality in predefined tasks. From the company's point of view, external solutions in particular can offer added value alongside in-house developments. Building on partnerships with software providers, such as Microsoft, companies like Siemens can benefit directly from new AI functionalities. This aligns with corporate goals, as the widespread adoption of external software solutions maximizes the reach of potential improvements. In the case of Microsoft, Siemens was able to evaluate the Microsoft Copilot at an early stage and make it available to selected employees as part of an internal pilot. This allowed the company to assess the Generative AI functionalities and build knowledge within the organization, laying the foundation for future implementation. Addressing the central research question, another key finding is the notable usability of Generative AI. Its ability to execute tasks based on natural language prompts offers a significantly streamlined user experience for interacting with AI systems. This, coupled with the accessibility of platforms like OpenAI's ChatGPT (and subsequent internal solutions), expands the scope of potential applications. For large corporations, this presents a significant shift, empowering employees without specialized AI knowledge to directly utilize the technology.

6.2. Emerging Patterns of Value Creation

Based on the case study results, the value creation potential of Generative AI has been outlined in Chapter 5. Use cases were clustered into broader schemes and key drivers were identified. Moreover, application areas were discovered which help companies like Siemens to structure and prioritize its use cases. With this, the case study supports closing the gap between scientific concepts and real-world application. It provides an answer to **Research Question 2**, namely in which ways Generative AI presents new avenues for value creation. Based on the technological innovation and usability as described above, the clustering of use cases in particular helps to understand the potential added value. Contrasted with the earlier work on value creation mechanisms, such as the process model of ML value creation by Shollo et al. (2022) or the four sources of value creation by Borges et al. (2021), similarities but also divergences become visible (see **Table** 4). Most noticeable, the use of AI for information retrieval can be seen as a new area for value creation which was not covered before. This can be traced back to the technological advances of Generative AI and its capability to process and provide answers from large data sets.

Another segment which was not touched on before, is coding support. While it can be argued that this could be seen as a subset of the cluster assistants, its far-reaching implications can serve as a reason for presenting it separately. With their knowledge of all major programming languages, Generative AI models can help to generate, adjust, test, and optimize code automatically, changing the way programmers deal with tasks and business problems. These unique capabilities of Generative AI could not be observed by researchers in

Table 4: Comparison of case study results with previous research

Case Study Clusters	**Shollo et al. (2022):** Shifting ML value creation mechanisms	**Borges et al. (2021):** The strategic use of artificial intelligence in the digital era
Assistants - Information Retrieval	N/A	N/A
Data Synthesis	**Knowledge Creation**	**Decision Support**
Assistants - Coaches & Helpers	**Task Augmentation**	**Customer and Employee Engagement**
Autonomous Assistants	**Autonomous Agents**	**Automation**
Content Creation	N/A	**New Products and Services**
Coding Support	N/A	N/A

2021 and 2022 respectively and therefore represent a valuable supplement to existing research.

6.3. Strategic and Operational Deployment

Besides the outlined avenues for value creation, supporting strategies and considerations have been identified. To address **Research Question 3**, which explores ways to harness Generative AI's full value creation potential, the following section discusses these findings in detail. Based on the insights gained regarding exploration and scaling, a new framework is proposed (**Figure 12**). It illustrates the connection between these phases, along with the necessary tasks to effectively exploit Generative AI's value creation potential. The hourglass shape symbolizes the critical interface between the two phases. It emphasizes the need for a business impact assessment to successfully scale use cases across the organization. This assessment requires weighing the costs and benefits for each use case individually. Since companies seek the best possible use of their limited resources, not all ideas from the exploration phase will necessarily be implemented. In order to determine the added value, subsequent questions should be addressed, namely who coordinates the implementation, how much development effort is required, which accompanying change management measures are necessary and who is responsible for the care, maintenance, and costs of the new solution.

The business impact assessment goes hand-in-hand with strategic considerations. For large companies, it is essential to view AI initiatives within a broader context and align them with corporate strategy (Kitsios & Kamariotou, 2021, p. 6). This is especially important since the added value of Generative AI use cases might not be immediately measurable. Companies should consciously define core areas to steer resources effectively. In this context, Generative AI presents a new challenge for managers and strategists. While past AI initiatives were more manageable and clearly definable, technological innovation now offers a myriad of potential value creation opportunities. Evaluating and filtering these, along with defining a clear strategy, becomes an essential task for successfully utilizing Generative AI. The framework suggests main tasks for each phase. In the exploration phase, these include creating awareness, sharing knowledge, building communities, and providing a common technology stack. To support the business impact assessment and strategic alignment, key tasks involve creating transparency, minimizing macrorisks, and allocating resources to identified use cases. Finally, to successfully scale solutions, companies should focus on harmonizing their data landscape, adjusting processes, and offering training and guidance to employees.

6.4. Theoretical Contributions

The results of this case study contribute to existing research in the fields of value creation, AI adoption, and capabilities. This thesis offers a valuable contribution by analyzing proposed frameworks by Borges et al. (2021) and Shollo et al. (2022) within the specific context of Generative AI. The analysis identifies areas where these models could be further expanded. Additionally, the framework by Uren and Edwards (2023), with its four lenses People, Processes, Technology, and Data proved highly applicable. This case study expands upon their model by detailing specific considerations within each category relevant to Generative AI implementation. Furthermore, new dimensions emerged for analysis, including the importance of timing, organizational setup, strategy, business, value and communication. The case study aligns with the findings of Caner and Bhatti (2020), highlighting the multiple perspectives from which AI can be viewed. For example, the results suggest that while Generative AI has overcome certain limitations like data labeling or generalizability of learning, others, such as biases and non-explainability, remain a challenge. Another interesting avenue for future research could be to investigate further application areas of Generative AI, as its primary use is currently in augmentation tasks, with autonomous systems seeming less prevalent. In conclusion, this thesis adds to the existing body of knowledge by refining and extending established models in the context of Generative AI. This detailed exploration offers valuable insights that can advance theoretical understanding and guide the development of more comprehensive frameworks for AI adoption.

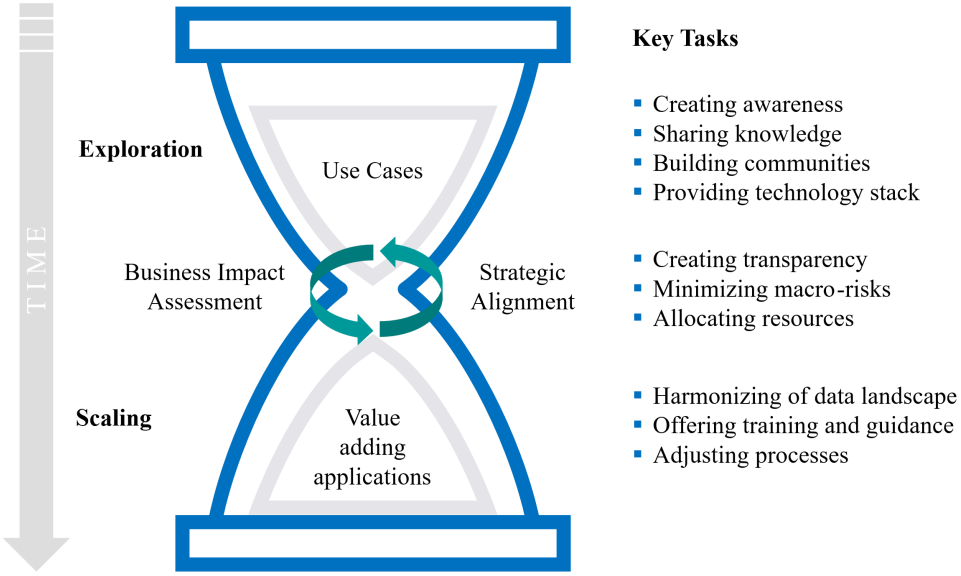


Figure 12: Proposed framework for exploration and scaling of Generative AI

6.5. Practical Implications

Merging the results from the case study and theoretical concepts on AI value creation and capabilities, practical implications emerge. In the following section, these should be outlined in more detail, providing companies with a clearer understanding of Generative AI and the required actions to create added value.

Firstly, companies should **create room for innovation**. Generative AI is unique in that its core functionalities (i.e., summarization, translation, and content creation) can be used by almost any employee. This empowers more people, even those without a background in IT or data science, to develop business ideas within their respective domains. Companies should embrace this opportunity and actively support the ideation and exploration of new Generative AI solutions.

An important prerequisite for this is the **provision of the required infrastructure**. With regard to Generative AI, this comprises company-internal platforms and software packages that employees can use safely and securely, without the risk of exposing sensitive company information externally. At the same time, it is important to highlight the limitations of the technology to avoid exaggerated expectations.

Secondly, companies should **fight for transparency.** Large corporations with their decentralized structures tend to develop similar solutions in multiple business areas. With Generative AI, this challenge becomes even more apparent as use cases can be developed within days or weeks. While exploration should be encouraged, creating transparency over all initiatives is key for companies; otherwise, resources risk being wasted on redundant solutions. Employees and departments should therefore be encouraged to share their knowledge and solutions with a broader audience. This should not be taken lightly: active management of potential use cases and clear guidance for employees are crucial to reduce isolated solutions, frictional losses, and unfulfilled expectations. If possible, top management should enforce the needed structures to provide a company-wide framework.

Thirdly, a **clear AI strategy** is needed to steer the company's resources in the right direction. As outlined before, Generative AI offers multifold opportunities for integration into processes and enhancement of existing products. However, companies should carefully evaluate each use case and its contribution to the overall company strategy. Regarding customer-facing solutions, the inherent technological risks need to be analyzed in detail. For internal processes, the added value of Generative AI must be weighed against increased complexity and consequential costs. To guide teams effectively, companies should provide guidance for their middle management, enabling them to evaluate new ideas within a broader context.

Finally, companies should focus on **bridging the gap between Generative AI knowledge and operational business units**. Since Generative AI use cases can increasingly be driven directly by these units without extensive IT or data science involvement, it is wise to empower them. Companies should consider strengthening existing departments (such as innovation, operational, or business excellence) or establishing dedicated Generative AI resources across the organization to maximize the technology's potential.

6.6. Study Limitations

Although providing valuable insights into the value creation opportunities of Generative AI, this case study is not without its limitations. The first limitation lies in the design of the study. As a single case study, it offers unique and comprehensive insights into the object of investigation; however, its findings cannot be broadly generalized. The second limitation is rooted in the author's active involvement in the exploration of Generative AI in the company. While

this allowed access to information that would normally not have been accessible (for example, the participation in external and internal conferences and the regular exchange with data scientists on technical aspects of Generative AI), biases and personal perceptions could impact the final results, despite the author's best efforts to avoid unintended interferences. The third limitation concerns the data used for the case study. Due to the novelty of Generative AI, existing concepts of AI value creation, adoption, and capabilities were used, concepts that might not fully reflect the latest technological advances. The selection of concepts was based on a structured approach; however, it cannot be ruled out that other unconsidered scientific work might have provided additional value. Regarding the data collected in the company, it should be noted that the selection of interview partners might have influenced the results. An attempt was made to cover as many areas and functions as possible, but a selection was necessary due to the substantial number of potential contacts.

In conclusion, this work represents a snapshot in time. The field of Generative AI is evolving rapidly. While the focus at the beginning of the case study was on textual understanding, new models now generate both images and videos. Despite the described limitations, this work aims to contribute to ongoing research on the value creation potential of Generative AI.

6.7. Future Research Opportunities

Building on the outlined limitations, future research could verify and expand the findings of this case study. Firstly, another case study in a similar setup could verify and strengthen the results, potentially adding to the proposed framework for exploration and scaling. Additionally, quantitative data would help deepen the understanding of a company's requirements and anticipated challenges when introducing Generative AI. Surveys among managers and employees could provide valuable insights into critical success factors. Moreover, more scientific work is needed to explain the unique capabilities of Generative AI compared to earlier machine learning models, and to integrate these findings into existing conceptual frameworks. Finally, future research should investigate the impact and value creation potential of the latest technological advances. With new models such as Google's Gemini, providing a context window of up to one million tokens, recently developed solutions for information retrieval could become obsolete. This and other advances such as OpenAI's text-to video model Sora continually open new avenues for research in this field.

7. Conclusion

In this study, the value creation opportunities of Generative AI were explored. Based on a single case study design, Siemens was chosen as the object of investigation as it offered a unique opportunity to study the phenomenon in depth. Over a period of one year, a considerable amount of time was spent observing and participating in numerous events to achieve a better understanding of the technology and Siemens' approach to introducing and exploiting its value creation potential. This was supplemented by 23 comprehensive interviews with key stakeholders, providing a complete view of use cases, activities, perspectives, and considerations.

The findings suggest that Generative AI presents new value creation opportunities, driven by technological advances and usability that bring AI closer to domain experts. Use cases have been identified across various areas and clustered into four main groups. Smart assistants with various levels of complexity form the majority of observed pilots, along with lighthouse projects in different business areas. While many applications offer potential added value, the study acknowledges that a considerable number of projects are still in their early stages.

Contributing to the ongoing discourse in AI research on value creation and required capabilities, different frameworks were applied to explain the case study findings. For example, addressing the capabilities for the successful implementation of AI, the outlined dimensions from previous research reappeared in the data. Other concepts could not be fully confirmed by the findings, opening room for further research. For example, existing work on value creation mechanisms was not able to explain all value creation clusters uncovered in the case study.

A key contribution of the thesis can be seen in its long-term observation of a company in the early adoption phase of a new technology. Starting in the early discovery stage only a few months after the release of Generative AI to the public, Siemens' steps to introduce, manage, and profit from the technology could be observed over a year. Building on the gathered knowledge, a new framework is proposed to explain the actions and strategies revealed in the case study. It introduces an exploration phase as an important time period in the adoption of Generative AI. Moreover, it identifies essential tasks to capitalize on the opportunities presented in this phase and outlines the necessary steps for a successful transition from exploration to scaling.

To conclude, the thesis provides practical implications and highlights potential areas for future research. Acknowledging its limitations, the thesis offers unique insights into a company's approach to Generative AI and contributes to the ongoing discourse on the potential value creation opportunities of the technology.

References

Akkiraju, R., Sinha, V., Xu, A., Mahmud, J., Gundecha, P., Liu, Z., Liu, X., & Schumacher, J. (2018). Characterizing machine learning process: A maturity framework. https://doi.org/10.48550/arXiv.1811.04871

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. https://doi.org/10.48550/arXiv.1409.0473

Bastian, M. (2023). GPT-4 has more than a trillion parameters. Retrieved March 25, 2023, from https://the-decoder.com/gpt-4-has-a-trillion-parameters/

Blanka, C., Krumay, B., & Rueckel, D. (2022). The interplay of digital transformation and employee competency: A design science approach. *Technological Forecasting and Social Change*, *178*, 121575. https://doi.org/10.1016/j.techfore.2022.121575

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., & Liang, P. (2021). On the Opportunities and Risks of Foundation Models. https://doi.org/10.48550/arXiv.2108.07258

Borges, A. F., Laurindo, F. J., Spínola, M. M., Gonçalves, R. F., & Mattos, C. A. (2021). The strategic use of artificial intelligence in the digital era: Systematic literature review and future research directions. *International Journal of Information Management*, *57*, 1–16. https://doi.org/10.1016/j.ijinfomgt.2020.102225

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brock, J. K.-U., & Wangenheim, F. v. (2019). Demystifying AI: What Digital Transformation Leaders Can Teach You about Realistic Artificial Intelligence. *California Management Review*, *61*(4), 110–134. https://doi.org/10.1177/1536504219865226

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., & Amodei, D. (2020). Language Models are Few-Shot Learners. https://doi.org/10.48550/arXiv.2005.14165

Brynjolfsson, E., Li, D., & Raymond, L. (2023). Generative AI at Work. https://doi.org/10.3386/w31161

Caner, S., & Bhatti, F. (2020). A Conceptual Framework on Defining Businesses Strategy for Artificial Intelligence. *Contemporary Management Research*, *16*(3), 175–206. https://doi.org/10.7903/cmr.19970

Davenport, T. H., & Ronanki, R. (2018). Artificial Intelligence for the Real World: Don't start with moon shots.

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2023). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. https://doi.org/10.2139/ssrn.4573321

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., & Wright, R. (2023). Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. http://arxiv.org/pdf/2303.10130v4

Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2023). Generative AI. *Business & Information Systems Engineering*. https://doi.org/10.1007/s12599-023-00834-7

Gartner. (2023). What Is Your Enterprise's AI Ambition? https://www.gartner.com/en/documents/4836831

Glaser, B. G., & Strauss, A. L. (1967). *The Discovery of Grounded Theory*. Routledge. https://doi.org/10.4324/9780203793206

Goldman Sachs. (2023). Generative AI could raise global GDP by 7%. https://www.goldmansachs.com/intelligence/pages/generative-ai-could-raise-global-gdp-by-7-percent.html

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org

Graves, A. (2012). *Supervised Sequence Labelling with Recurrent Neural Networks* (Vol. 385). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-24797-2

Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for boltzmann machines. *Cognitive Science*, *9*(1), 147–169. https://doi.org/10.1016/S0364-0213(85)80012-4

Hochreiter, S., Bengio, Y., Frasconi, P., & Schmidhuber, J. (2001). Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. *A field guide to dynamical recurrent neural networks*. https://doi.org/10.1109/ICASSP.2001.940859

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Kim, Y., Denton, C., Hoang, L., & Rush, A. M. (2017). Structured Attention Networks. https://doi.org/10.48550/arXiv.1702.00887

Kitsios, F., & Kamariotou, M. (2021). Artificial Intelligence and Business Strategy towards Digital Transformation: A Research Agenda. *Sustainability*, *13*(4), 2025. https://doi.org/10.3390/su13042025

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lee, J., Suh, T., Roy, D., & Baucus, M. (2019). Emerging Technology and Business Model Innovation: The Case of Artificial Intelligence. *Journal of Open Innovation: Technology, Market, and Complexity*, *5*(3), 44. https://doi.org/10.3390/joitmc5030044

Lichtenthaler, U. (2019). An Intelligence-Based View of Firm Performance: Profiting from Artificial Intelligence. *Journal of Innovation Management*, *7*(1), 7–20. https://doi.org/10.24840/2183-0606_007.001_0002

Luger, G. F., & Stubblefield, W. A. (1998). *Symbolic Computation* (3rd). Addison-Wesley.

McCarthy, J. (1958). Programs with common sense: Proceedings of the Symposium on Mechanisation. London77-84.

McCarthy, J., & Feigenbaum, E. A. (1990). In Memoriam: Arthur Samuel: Pioneer in Machine Learning. *AI Magazine*, *11*(3), 10. https://doi.org/10.1609/aimag.v11i3.840

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

Myers, M. D. (2020). *Qualitative Research in Business & Management* (Third). SAGE Publications Ltd.

Ng, A. Y., & Jordan, M. I. (2001). On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes.

Nilsson, N. J. (2013). *The Quest for Artificial Intelligence*. Cambridge University Press. https://doi.org/10.1017/CBO9780511819346

OpenAI. (2015). Introducing OpenAI. https://openai.com/blog/introducing-openai

OpenAI. (2022). Introducing ChatGPT. https://openai.com/blog/chatgpt

OpenAI. (2023). GPT-4 Technical Report. http://arxiv.org/pdf/2303.08774v3

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. http://arxiv.org/pdf/2203.02155v1

Peng, B., Alcaide, E., Anthony, Q., Albalak, A., Arcadinho, S., Biderman, S., Cao, H., Cheng, X., Chung, M., Grella, M., GV, K. K., He, X., Hou, H., Lin, J., Kazienko, P., Kocon, J., Kong, J., Koptyra, B., Lau, H., & Zhu, R.-J. (2023). RWKV: Reinventing RNNs for the Transformer Era. https://doi.org/10.48550/arXiv.2305.13048

Radford, A., & Narasimhan, K. (2018). Improving Language Understanding by Generative Pre-Training.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. https://api.semanticscholar.org/CorpusID:160025533

Rao, A. (2017). A Strategist's Guide to Artificial Intelligence. https://www.strategy-business.com/article/A-Strategists-Guide-to-Artificial-Intelligence

Reilly, A., Depa, J., & Douglass, G. (2019). AI: Build to Scale: From Experimental to Exponential. https://www.accenture.com/content/dam/accenture/final/a-com-migration/thought-leadership-assets/accenture-built-to-scale-pdf-report.pdf

Reim, W., Åström, J., & Eriksson, O. (2020). Implementation of Artificial Intelligence (AI): A Roadmap for Business Model Innovation. *AI*, *1*(2), 180–191. https://doi.org/10.3390/ai1020011

Rubin, H. J., & Rubin, I. S. (2012). *Qualitative interviewing: The art of hearing data*. SAGE Publications.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0

Russell, S. J., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach* (Fourth). Pearson.

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, *3*(3), 535–554. https://doi.org/10.1147/rd.33.0210

Shollo, A., Hopf, K., Thiess, T., & Müller, O. (2022). Shifting ML value creation mechanisms: A process model of ML value creation. *The Journal of Strategic Information Systems*, *31*(3), 1–20. https://doi.org/10.1016/j.jsis.2022.101734

Siemens. (2023). Annual Report 2023. https://www.siemens.com/global/en/company/investor-relations/events-publications-ad-hoc/annualreports.html

Siemens. (2024). Siemens - Who we are. https://www.siemens.com

Simmons, A. B., & Chappell, S. G. (1988). Artificial intelligence-definition and practice. *IEEE Journal of Oceanic Engineering*, *13*(2), 14–42. https://doi.org/10.1109/48.551

Sutton, R. S., & Barto, A. (2018). *Reinforcement learning: An introduction* (Second edition). The MIT Press.

Trunk, A., Birkel, H., & Hartmann, E. (2020). On the current state of combining human and artificial intelligence for strategic organizational decision making. *Business Research*, *13*(3), 875–919. https://doi.org/10.1007/s40685-020-00133-x

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, *59*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Uren, V., & Edwards, J. S. (2023). Technology readiness and the organizational journey towards AI adoption: An empirical study. *International Journal of Information Management*, *68*, 102588. https://doi.org/10.1016/j.ijinfomgt.2022.102588

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. http://arxiv.org/pdf/1706.03762v5

Voita, E., Sennrich, R., & Titov, I. (2019). The Bottom-up Evolution of Representations in the Transformer: A Study with Machine Translation and Language Modeling Objectives. https://doi.org/10.48550/arXiv.1909.01380

Weber, M., Engert, M., Schaffer, N., Weking, J., & Krcmar, H. (2023). Organizational Capabilities for AI Implementation—Coping with Inscrutability and Data Dependency in AI. *Information Systems Frontiers*, *25*(4), 1549–1569. https://doi.org/10.1007/s10796-022-10297-y

Yin, R. K. (2018). *Case study research and applications: design and methods* (6th ed.). SAGE.

Ziegler, D. M., Stiennon, N., Wu, J. [, Brown, T. B., Radford, A., Amodei, D., Christiano, P., & Irving, G. (2019). Fine-Tuning Language Models from Human Preferences. https://doi.org/10.48550/arXiv.1909.08593

Junior Management Science 10(3) (2025) 657-676

Junior Management Science

www.jums.academy
ISSN: 2942-1861



The Effect of Changes in Internal Control Systems on Audit Risk

Justus Olbrich

Ludwig Maximilian University of Munich

Abstract

Internal control weaknesses influence audit fees and audit risk, making their remediation a crucial aspect of corporate governance. While prior research focuses on auditors, this study examines the corporate perspective, analyzing how the remediation of internal control weaknesses affects audit fees and audit risk. Using a dataset of 450 observations, the audit fee model shows a significant negative association between audit fees and the remediation of internal control weaknesses, indicating that audit fees decrease as the company remediates its internal control weaknesses. The restatement model shows no significant relationship between abnormal audit fees and the probability of restatements, suggesting that changes in audit fees due to changes in internal control systems have no impact on audit risk. However, the remediation of internal control weaknesses is significantly negatively associated with the probability of restatements, meaning that firms that remediate their internal control weaknesses experience a lower audit risk and higher audit quality. These findings highlight the economic significance of internal control quality and its implications for firms, regulators, and auditors in mitigating audit risk.

Keywords: audit fee model; audit quality; audit risk model; internal control weaknesses; restatement model

1. Introduction

As a consequence of accounting scandals at Enron, WorldCom, and Tyco in the early 2000s, the United States Congress enacted the Sarbanes-Oxley Act of 2002 (SOX).[1] The U.S. government's goal with SOX is to regain public confidence and information quality by improving the accuracy of reported financial statements.[2] With the introduction of SOX, researchers and other stakeholders could evaluate approximately the quality of a firm's internal control systems.[3] In more detail, mandatory disclosure requirements for audit fees and internal control deficiencies make it possible to examine audit risk and its components.[4] However, SOX brings not only advantages. For example, companies must record additional costs in case of a negative report on internal control systems.[5] Furthermore, there is evidence that the implementation of SOX and the growing complexity of business transactions are causing increases in the number of restatements overall and the number of firms reporting a restatement.[6] These increases are mainly due to adjustments to revenues, costs, or expenses in the financial statements.[7] Specifically, the number of annual announcements of financial restatements increased by 67 percent between 2002 and 2005.[8] Also, with a more recent reference, the Public Com-

I would like to express my sincere gratitude to Prof. Dr. Christian Hofmann, Sebastian Kuhn, and the Institute for Accounting and Control at Ludwig-Maximilians-Universität München for their support in completing this thesis. I particularly appreciated the welcoming and academically stimulating atmosphere at the institute. The guidance and constructive feedback provided by Sebastian Kuhn were instrumental to the success of this research. His availability and valuable input significantly shaped the final outcome of this work.

[1] See therefore Guragai and Hutchinson (2019, p. 363); Hoag and Hollingsworth (2011, p. 173); Raghunandan and Rama (2006, p. 99).

[2] See therefore Ashbaugh-Skaife et al. (2008, p. 219); Guragai and Hutchinson (2019, p. 362); Hoag and Hollingsworth (2011, p. 174).

[3] See therefore Chalmers et al. (2019, p. 100).

[4] See therefore Seidel (2017, p. 1345).

[5] See therefore Munsif et al. (2011, p. 87).

[6] See therefore DeFond and Zhang (2014, p. 275); Kester et al. (2013, pp. 155-156).

[7] See therefore Kester et al. (2013, p. 156).

[8] See therefore United States Government Accountability Office (2006, p. 11).

DOI: https://doi.org/10.5282/jums/v10i3pp657-676

© The Author(s) 2025. Published by Junior Management Science.
This is an Open Access article distributed under the terms of the CC-BY-4.0 (Attribution 4.0 International). Open Access funding provided by ZBW.

pany Accounting Oversight Board reported an alert due to a significant number of financial reporting control deficiencies in 2013.[9] This development underlines the interest of several stakeholders, like regulators, audit firms, and corporate boards.[10] For the latter, the consequences of internal control weaknesses and financial statement restatements are particularly significant, as stated in the following.

Internal control weaknesses can impact the firm itself and the firm's environment. On the one hand, internal control quality can impact decision-making and the actions taken by the management due to the reliability of forecasts.[11] For example, Chalmers et al. (2019) point out that weaknesses in internal control cause errors in the internal management report, which ultimately lead to uncertainty and lower accuracy of earnings forecasts and thus to changes in behavior.[12] On the other hand, as indicated above, adverse disclosures on the internal control system can impact the firm economically.[13] Negative stock price reactions, as well as the limited access to funds, are exemplary consequences of internal control weakness disclosures.[14] Among other things, this lies in the relationship between investors and the company. This relationship is affected since internal control weaknesses represent a relevant risk for investors.[15] Investors' risk is that the statements are not presented fairly and are therefore not credible.[16] In addition, credit-rating agencies like Moody's or Fitch Ratings have suggested that internal control weaknesses lead to less reliable financial statements.[17] Consequently, investors take advantage of internal control disclosures to value and allocate capital.[18] Therefore, companies and their managers have an incentive to reduce the risk of a misreporting financial statement by remediating internal control weaknesses so that creditors and other stakeholders fairly state their firms.[19]

Financial restatements can also negatively impact the firm and its value. This applies especially when the restated financial report is worse than the original.[20] Restatements indicate low reporting quality and greater information risk like internal control weaknesses.[21] Following that, restatements indicate weak reliability of a company's financial reporting system and therefore lead to a negative image for the firm.[22] Consequently, analysts have a lower forecast optimism and reduce their stock suggestions quicker.[23] As evidence, the Securities and Exchange Commission (SEC) lists financial restatements as a significant factor undermining investors' confidence in financial reporting.[24] This lower optimism and confidence causes that, in the end, restatement firms have significant negative stock returns and higher costs of capital.[25]

In summary, the consequences of internal control weaknesses and financial statement restatements are particularly significant for the companies due to the economic impact. Moreover, consequences exist not only when internal control weaknesses arise but also when internal control weaknesses remediate. Thus, examining these remediations effects from a corporate perspective is interesting because they can specifically try to eliminate their internal control weaknesses. Therefore, I want to examine how changes in internal control systems due to remediation of internal control weaknesses affect audit risk, respectively, the probability of restatements.

This study contributes to prior literature because many studies interpret their results from an auditor's perspective or as recommendations for regulators.[26] Whereas in this study, the focus is on the company. Thus, the study results interpret in a corporate sense.

As Hogan and Wilkins (2008) or Blankely et al. (2012) used, I first use an Audit Fee Model to determine the relationship between the remediation of internal control weaknesses and audit fees.[27] In the next step, I use a Restatement Model to examine the effect of abnormal audit fees and the remediation of internal control weaknesses on the overall audit risk.[28] Here, the probability of a financial statement restatement acts as a proxy for the audit risk.

This paper is structured as follows: The following section discusses the previous literature and presents the hypotheses. The subsequent section discusses the data, research methods, and results. The final section discusses interpretations and conclusions.

2. Theoretical Analysis and Derivation of Hypotheses

As mentioned above, section 404 of the SOX requires publicly traded companies to disclose a management and auditor report about the effectiveness of their internal control systems.[29] The disclosures are in the form of 10-K reports. According to the Committee of Sponsoring Organizations of the Treadway Commission (COSO), internal control is "a process, effected by an entity's board of directors, management, and

[9] See therefore Lawrence et al. (2018, p. 140); PCAOB (2013, p. 35).
[10] See therefore Romanus et al. (2008, p. 389).
[11] See therefore Chalmers et al. (2019, p. 95).
[12] See therefore Chalmers et al. (2019, p. 95).
[13] See therefore Albring et al. (2018, p. 485).
[14] See therefore Hammersley et al. (2012, p. 74); Hammersley et al. (2008, p. 141).
[15] See therefore Hoitash et al. (2008, p. 106).
[16] See therefore Bae et al. (2021, p. 587); DeFond and Zhang (2014, p. 276).
[17] See therefore Raghunandan and Rama (2006, p. 102).
[18] See therefore Hammersley et al. (2012, p. 74); Jonas et al. (2007, p. 1).
[19] See therefore Mitra et al. (2017, p. 240).
[20] See therefore Pittman and Zhao (2021, p. 129).
[21] See therefore Pittman and Zhao (2021, p. 129).
[22] See therefore Feldmann et al. (2009, p. 209); Kester et al. (2013, p. 155).
[23] See therefore Cao et al. (2012, p. 958); Fang and Yasuda (2009, p. 3735); Jackson (2005, p. 706).
[24] See therefore Archambeault et al. (2008, p. 966); Guragai and Hutchinson (2019, p. 364).
[25] See therefore H. Chen et al. (2011, p. 2); Hribar and Jenkins (2004, p. 337); Palmrose et al. (2004, p. 59).
[26] See therefore Ashbaugh-Skaife et al. (2008, p. 217); Elder et al. (2009, p. 543); Krishnan et al. (2011, p. 1).
[27] See therefore Hogan and Wilkins (2008, p. 227); Blankely et al. (2012, p. 83).
[28] See therefore Blankely et al. (2012, p. 84).
[29] See therefore Krishnan et al. (2011, p. 1).

other personnel, designed to provide reasonable assurance regarding the achievement of objectives relating to operations, reporting, and compliance."[30]

As part of the enactment of SOX, the Public Company Accounting Oversight Board (PCAOB) emerged. The PCAOB is a private non-profit organization that oversees the audits of public companies.[31] Therefore, the PCAOB established and amended auditing and quality control standards to fulfill its objectives. In the following, I refer to three of these auditing standards: Auditing Standard No. 5, Auditing Standard No. 8, and Auditing Standard No. 13.

Auditing Standard No. 5 describes a top-down approach consisting of five phases of the internal control over financial reporting (ICOFR) audit: planning, scoping, testing, evaluation, and reporting.[32] This approach aims to make ICOFR audits more efficient and, therefore, more economical for the companies.[33] The PCAOB defines ICOFR as "a process [...] to provide reasonable assurance regarding the reliability of financial reporting and the preparation of financial statements for external purposes in accordance with GAAP."[34] Within the successive phases of this approach, the auditor should identify and evaluate areas that can lead to a material weakness.[35] A material weakness relates to "a deficiency [...] in internal control over financial reporting, such that there is a reasonable possibility that a material misstatement of the company's annual or interim financial statements will not be prevented or detected on a timely basis."[36] Consequently, the internal control system is considered ineffective due to one or more material weaknesses.[37] Moreover, the PCAOB distinguishes between material weaknesses and significant deficiencies. The latter ones are "deficienc[ies] [...] in internal control over financial reporting that [are] less severe than material weaknesses".[38] Nevertheless, prior research suggests distinguishing between significant deficiencies and material weaknesses is complex and challenging.[39] Moreover, there is no consensus in the literature concerning the impact of problem severity on the relationship between audit fees and internal control weaknesses. For example, Hoitash et al. (2008) compare the impact of material weaknesses and significant deficiencies on audit fees and find that only material weaknesses are significantly associated.[40] Whereas Raghunandan and Rama (2006) report, that audit pricing does not vary depending on the severity of material weaknesses.[41] These results support the condition in this study of not distinguishing between material weaknesses and significant deficiencies, respectively, the severity of deficiencies.

In summary, the objective of Auditing Standard No. 5 is to conduct the audit of financial statements in a reliable manner that reduces audit risk to an appropriately low level. Therefore, the PCAOB additionally includes the Audit Risk Model in their Auditing Standard No. 8.[42]

The Audit Risk Model is a framework for evaluating risks in the audit planning process.[43] This conceptual tool decomposes audit risk into inherent risk, control risk, and detection risk.[44]

Audit Risk Model:

$$\text{Audit Risk} = \text{Inherent Risk} * \text{Control Risk} * \text{Detection Risk}$$

Inherent risk "refers to the susceptibility of an assertion to a misstatement, due to error or fraud, that could be material, individually or in combination with other misstatements, before consideration of any related controls."[45] Whereas control risk refers to "the risk that a misstatement due to error or fraud [...] will not be prevented or detected on a timely basis by the company's internal control."[46] Inherent and control risk build together the risk of material misstatement.[47] In other words, they represent the risk that is coming from the company itself. In contrast, detection risk refers to "the risk that the procedures performed by the auditor will not detect a misstatement that exists and that could be material, individual or in combination with other misstatements."[48] The three risk components together form the audit risk which is "the risk that the auditor expresses an inappropriate audit opinion when the financial statements are materially misstated."[49] Closely related to the Audit Risk Model in Auditing Standard No. 8 is Auditing Standard No. 13. Its objective is that auditors "address the risks of material misstatement through appropriate overall audit responses and audit procedures."[50] In other words, this regulation requires that auditors adjust their effort put into audit to the previously assessed risk.[51] Regarding the Audit Risk Model, this would mean that if, for example, control risk increases, auditors would increase their audit effort to decrease detection risk and therefore keep audit risk at an appropriate level. Conversely, auditors will reduce their effort if the control risk decreases.

Several academic studies have examined this topic, given this regulatory emphasis and the importance of the Audit

[30] Committee of Sponsoring Organizations of the Treadway Comission (2012, p. 1).
[31] See therefore Elder et al. (2009, p. 544).
[32] See therefore Asare et al. (2013, pp. 3-4); Krishnan et al. (2011, p. 2); PCAOB (2007, p. 18).
[33] See therefore Hoag and Hollingsworth (2011, p. 176); Krishnan et al. (2011, p. 2).
[34] PCAOB (2007, pp. 57-58).
[35] See therefore Asare et al. (2013, p. 5).
[36] PCAOB (2007, p. 59).
[37] See therefore PCAOB (2007, p. 20).
[38] PCAOB (2007, p. 60).
[39] See therefore Hoitash et al. (2008, p. 123).
[40] See therefore Hoitash et al. (2008, p. 118).
[41] See therefore Raghunandan and Rama (2006, p. 99).
[42] See therefore PCAOB (2010, pp. 11-12).
[43] See therefore R. S. Chen (2019, p. 99).
[44] See therefore Elder et al. (2009, p. 544).
[45] PCAOB (2010, p. 11).
[46] PCAOB (2010, p. 12).
[47] See therefore PCAOB (2010, pp. 11-12).
[48] PCAOB (2010, p. 12).
[49] PCAOB (2010, p. 11).
[50] PCAOB (2010, p. 64).
[51] See therefore Feldmann et al. (2009, p. 207); Pittman and Zhao (2021, p. 130).

Risk Model and its components. For example, one set of studies examines the characteristics of companies reporting an internal control weakness, leading to higher control risk. Firms reporting internal control deficiencies tend to be "smaller, younger, financially weaker, more complex, growing rapidly, or undergoing restructuring."[52] Additionally, Y. Chen et al. (2014) report that information technology capabilities support internal control systems and signal lower business risk, leading to lower audit fees.[53]

Other studies, in turn, examine the relationship between audit effort, respectively audit fees, and internal control deficiencies. On the one hand, Albring et al. (2018) examine the impact of audit fees on internal control weaknesses. They suggest that prior year audit fees can function as a signal for future internal control weaknesses.[54] The explanation for this result is that internal control weaknesses in the previous year increase risk but do not yet warrant reporting. Nevertheless, audit fees increase due to the increased risk. In the following year, it changes, and the reporting of the internal control weaknesses occurs.[55] On the other hand, prior literature examines the impact of internal control weaknesses on audit fees. In general, empirical evidence supports the previously mentioned theoretical view of the Audit Risk Model that weaknesses in internal control are associated with higher audit effort and, therefore, higher audit fees.[56] Elder et al. (2009) also find that higher audit fees are associated with internal control weaknesses. Additionally, they suggest a pecking order among auditors' strategies to manage an increase in control risk.[57] Thus, at first, auditors adjust their fees, then modify their opinions, and at last resign.[58]

Another topic within this field of research is the effect of remediation of material weaknesses. In line with the literature and theoretical considerations of the Audit Risk Model, Krishnan et al. (2011) find that the remediation of internal control weaknesses is associated with a decrease in audit fees.[59] Following that, Hammersley et al. (2012) report that companies unable to remediate their material weaknesses notice higher audit fees and a higher likelihood of auditor resignation.[60] In addition to these findings, some suggest that companies continue to pay significantly higher audit fees even after the material weakness remediation was successful.[61] For example, Munsif et al. (2011) find audit fee risk premiums of 35% one year and 21% two years after the remediation of internal control weaknesses.[62] As an explanation, Calderon et al. (2012) state that the amount of audit fees determines within the initial risk assessment. Further, this initial risk assessment strongly considers audit results of prior periods.[63] A different explanation comes from Albring et al. (2012), who assert that auditors do not reduce their effort in the remediation year, which ultimately remains audit fees high.[64] From these explanations and results, the question arises whether audit effort or a risk premium are primary drivers for audit fees.

Several studies examine whether the change in audit fees results from a change in effort or a change in the risk premium. The theoretical foundation for this consideration lies in Simunic's (1980) audit pricing model.[65] This model suggests that audit efforts do not solely drive audit fees. There is also an expected future loss cost component representing litigation risk.[66] In other words, an increase in control risk through an internal control weakness leads to an increase in possible auditor litigation. Moreover, there is no possible reduction in litigation risk, and the auditor charges a risk premium.[67] Some previous literature supports the audit pricing model. Jiang and Son (2015) find that auditors consider adjustments in audit effort and risk premium in response to increasing control risk.[68] Bae et al. (2021) can confirm these results and further state that, next to a significantly higher audit effort, companies pay a 6 percent premium in response to internal control weaknesses.[69] Hoag and Hollingsworth (2011) report a 42 percent audit fee premium for Big4 clients.[70]

In summary, the literature supports the view that internal control weaknesses are positively associated with audit fees. This association also applies to the remediation of internal control weaknesses as it decreases audit fees. Further, there is a positive relationship between audit fees and audit effort. In addition to the effort component within the audit fee, there seems to be a risk premium component. In the context of this study, I conclude that a change in audit effort at least partially causes an audit fee change. Therefore, the first hypothesis considers this conclusion and states:

> H1: The remediation of an internal control weakness is associated with a decrease in audit fees for that firm.

As the Audit Risk Model indicates, the change of effort respectively the change in audit fees can affect the audit risk (PCAOB Auditing Standard No. 8 §3). However, previous literature often considers audit quality instead of audit risk. Furthermore, a negative relationship between audit risk and audit quality is assumed. This negative association is derived

[52] Doyle et al. (2007, p. 193).
[53] See therefore Y. Chen et al. (2014, p. 175).
[54] See therefore Albring et al. (2018, p. 499).
[55] See therefore Albring et al. (2018, p. 488).
[56] See therefore Bae et al. (2021, p. 616); Gaber et al. (2019, p. 16); Hay (2013, p. 163); Seidel (2017, p. 1342).
[57] See therefore Elder et al. (2009, p. 545).
[58] See therefore Elder et al. (2009, p. 543).
[59] See therefore Krishnan et al. (2011, p. 1).
[60] See therefore Hammersley et al. (2012, p. 73).
[61] See therefore Calderon et al. (2012, p. 693); Mitra (2009, p. 372).
[62] See therefore Munsif et al. (2011, p. 87).
[63] See therefore Calderon et al. (2012, p. 696).
[64] See therefore Albring et al. (2012, pp. 377, 380).
[65] See therefore Simunic (1980, p. 169).
[66] See therefore R. S. Chen (2019, p. 99); Jiang and Son (2015, p. 321).
[67] See therefore Jiang and Son (2015, p. 321).
[68] See therefore Jiang and Son (2015, p. 318).
[69] See therefore Bae et al. (2021, p. 616).
[70] See therefore Hoag and Hollingsworth (2011, p. 196).

from the definition of audit risk as it is "the risk that the auditor expresses an inappropriate audit opinion when the financial statements are materially misstated."[71] In other words, an increase in audit risk is associated with a deterioration of audit quality. Therefore, I examine the relationship between audit fees and audit quality in the following.

High audit quality represents the "assurance that financial statements faithfully reflect the firm's underlying economics, conditioned on its financial reporting system and innate characteristics."[72] This definition includes auditor error detection as well as compliance with auditing standards.[73] Instead of total audit fees, the literature considers abnormal audit fees since "normal fees are mainly determined by factors that are common across different clients [...], while abnormal fees are determined by factors that are idiosyncratic to a specific auditor-client relationship."[74] Abnormal audit fees are the "difference between actual audit fees [...] and the expected, normal level of audit fees."[75] There are two contradictory approaches in the literature, the effort, and the economic bonding view, explaining the relationship between abnormal audit fees and audit quality.

The effort view refers to the view of the PCAOB in Auditing Standard No. 13. Theoretically, it describes that more effort is associated with higher audit fees, and due to more audit effort, the audit quality increases.[76] Additionally, the effort view assumes that increased or reduced audit effort causes abnormal audit fees.[77] Therefore, in other words, abnormal audit fees are positively associated with audit quality. Empirically, there is some evidence for the effort view. Eshleman and Guo (2014) find a negative relationship between the abnormal audit fees and the likelihood that companies use income-increasing discretionary accruals. This result implies that companies and their managers lose discretion in choosing accruals by increasing the abnormal audit fees.[78] Moreover, this result suggests a positive relationship between abnormal audit fees and audit quality which eventually supports the effort view.[79] The findings from Blankely et al. (2013) also align with the effort view. Here, they examine the effect of abnormal audit fees on future restatements and find a negative relationship. [80] Therefore, they conclude that lower abnormal audit fees indicate lower audit effort, resulting in a higher likelihood of restatements, respectively, in lower audit quality.[81] Lobo and Zhao (2013) support this view by finding a robust negative association between audit efforts and restatements.[82] In addition, they argue that the discrepant results are due because pre-audit misstatement risk is not constant.[83] Alhadab (2018) finds supporting results but uses a different explanatory approach.[84] He argues by using the productive theory, which states that additional audit fees lead to additional audit procedures by the audit firm, which ultimately increases the efficiency and effectiveness of the companies' operating system. This increase in the efficiency and effectiveness of the companies' operating system leads to less ability for earnings manipulation by managers and, therefore, to higher audit quality and higher financial reporting quality.[85]

Economic bonding stems from the theoretical consideration that high audit fees bond the auditor to its client, the company. Due to this bond, the auditor loses his or her independence and, therefore, also his or her professional skepticism, which ultimately results in auditors giving companies more discretion in questionable accounting practices.[86] According to the economic bonding view, the auditor's independence to the entity determines the ability to detect a misstatement and not the effort the auditor exerts.[87] In terms of the association between abnormal audit fees and audit quality, this would mean that there is a negative relationship between them. Several papers support the economic bonding view and provide empirical evidence. Hribar et al. (2014) find a significant positive relationship between unexplained audit fees, which equal abnormal audit fees, and the occurrence of restatements.[88] Here, it is worth mentioning that they reverse the direction of correlation and therefore argue that high audit fees result from a low-quality accounting system.[89] Despite this, they cannot rule out the economic bonding effect between poor quality accounting and fees.[90] Hoitash et al. (2007) report a negative relation between total fees and audit quality, suggesting that higher fees increase the economic dependence of auditors.[91] There is also evidence explicitly rejecting the economic bonding view. Blankely et al. (2014) state that restating firms have longer abnormal audit report lags and thus a more significant number of days between the fiscal year-end and the audit report date.[92] They argue that there should be a negative relationship between future restatements and audit report lags in the case of economic bonding.[93] The idea is that a less skeptical auditor, due to his or her missing independence, fails to adjust his or her effort, resulting in a shorter audit report lag. Literature also differentiates between positive and negative abnormal audit fees when considering economic bonding effects. Choi et al.

[71] PCAOB (2010, p. 11).
[72] DeFond and Zhang (2014, p. 276).
[73] See therefore Rajgopal et al. (2021, p. 560).
[74] Choi et al. (2010, p. 116).
[75] Choi et al. (2010, p. 116).
[76] See therefore Eshleman and Guo (2014, p. 120).
[77] See therefore Blankely et al. (2013, p. 18); Eshleman and Guo (2014, p. 135).
[78] See therefore Eshleman and Guo (2014, p. 129).
[79] See therefore Eshleman and Guo (2014, p. 129).
[80] See therefore Blankely et al. (2013, p. 18).
[81] See therefore Blankely et al. (2013, p. 18).
[82] See therefore Lobo and Zhao (2013, p. 1385).
[83] See therefore Lobo and Zhao (2013, p. 1388).
[84] See therefore Alhadab (2018, p. 395).
[85] See therefore Alhadab (2018, pp. 395, 397).
[86] See therefore Blankely et al. (2014, p. 30); Eshleman and Guo (2014, p. 120).
[87] See therefore Blankely et al. (2013, p. 16); Hoitash et al. (2007, p. 762).
[88] See therefore Hribar et al. (2014, pp. 516, 517, 528).
[89] See therefore Hribar et al. (2014, p. 514).
[90] See therefore Hribar et al. (2014, p. 535).
[91] See therefore Hoitash et al. (2007, pp. 761-762).
[92] See therefore Blankely et al. (2014, p. 27).
[93] See therefore Blankely et al. (2014, pp. 27, 30).

(2010) and Krauß et al. (2015) find only a significant negative association between audit quality and positive abnormal audit fees, consistent with the economic bonding view.[94] Negative abnormal fees do not affect the audit quality. They argue that auditors do not compromise audit quality since they have no monetary incentives.[95] At the same time, audit quality deteriorates if the monetary benefits exceed the associated costs for a misstated audit report.[96] Different results for negative abnormal audit fees are in the paper of Asthana and Boone (2012). They report a negative relationship between audit quality and negative abnormal audit fees. They argue that companies with low audit fees have greater bargaining power, giving them more freedom for questionable accounting practices.[97]

In summary, there is no clear relation between abnormal audit fees and audit quality. Neither the effort nor economic bonding view is a generally prevailing explanatory approach. However, the effort view and its assumption better suit the context of the first hypothesis. Thus, I assume that increased or reduced audit effort causes abnormal audit fees.[98] Accordingly, in the context of this study and related to hypothesis 1, I expect lower audit fees due to the remediation of internal control weaknesses, implying reduced effort and thus poorer audit quality, respectively higher audit risk.[99] Finally, this suggests a negative association between abnormal audit fees and audit risk, which represents the second hypothesis:

> H2a: Abnormal audit fees are negatively associated with audit risk.

In addition to the indirect examination of the effects of changes in internal control systems on audit risk via audit fees, I examine the direct impact of remediating an internal control weakness on audit risk. Therefore, I expect that the remediation of an internal control weakness is associated with lower audit risk. The last hypothesis is:

> H2b: The remediation of an internal control weakness is associated with a decrease in audit risk.

3. Empirical Analysis of the Effect of Changes in Internal Control Systems on Audit Risk

3.1. Research Methods

In the following, I present the two regression models examining the hypotheses. The first one, for hypothesis 1, is an Audit Fee Model to examine the relationship between the remediation of internal control weaknesses and audit fees. For hypotheses 2a and 2b, the second one is a Restatement Model to examine the effect of remediating internal control weaknesses and abnormal audit fees on audit risk, respectively, the probability of a restatement. In addition, this chapter considers the sample attrition and composition.

3.1.1. Audit Fee Model

The first hypothesis addresses the relationship between the remediation of internal control weaknesses and audit fees. In constructing the model, I rely on previous models concerning audit fees. In these Audit Fee Models, the dependent variable is the natural log of audit fees (*LNAFEE*).[100] The explanatory variable is the occurrence of remediation of internal control weaknesses (ΔICW). This variable will eventually allow me to examine the effect of remediation in internal control systems on audit fees. Therefore, I initially test the first hypothesis using the following regression model:

Audit Fee Model:

$$\begin{aligned} LNAFEE_{i,t} = {} & \beta_0 + \beta_1 \Delta ICW_{i,t} + \beta_2 LNTA_{i,t} \\ & + \beta_3 LNSALES_{i,t} + \beta_4 MERGER_{i,t} \\ & + \beta_5 INVREC_{i,t} + \beta_6 CA_{i,t} + \beta_7 INTANG_{i,t} \\ & + \beta_8 ROA_{i,t} + \beta_9 LOSS_{i,t} + \beta_{10} CR_{i,t} \\ & + \beta_{11} LEV_{i,t} + \beta_{12} BIG4 + \beta_{13} BUSY_{i,t} \\ & + \beta_{14} OPINON_{i,t} + \beta_{15-23} INDCON_{i,t} \\ & + \beta_{24-33} YEAR_{i,t} + \varepsilon_{i,t} \end{aligned}$$

The primary variable of interest (ΔICW) concerns changes in internal control systems. Instead of considering the simple occurrence of internal control weaknesses, I consider the remediation of them because, in the first case, the previous level of weaknesses is unknown. For example, it could be possible that companies have internal control weaknesses but do not have to disclose them because their occurrence is below a particular threshold.[101] Another scenario would be if the company remediated the internal control weakness before disclosing it.[102] Thus, it would be unclear to which extent the internal control system has changed. Moreover, the effect of the occurrence of internal control weaknesses would be biased.[103] In the latter case, the disclosure of internal control weakness makes known the initial level of internal control weaknesses. Therefore, I use ΔICW as a dummy variable that takes the value of 1 if a firm reports an internal control weakness in $t = 0$ but does not report this weakness in $t = 1$. In other words, the dummy variable takes the value of 1 if a firm remediates its internal control weakness. Consistent with previous studies, I expect a negative relationship between *LNAFEE* and ΔICW, indicating that

[94] See therefore Choi et al. (2010, p. 116); Krauß et al. (2015, pp. 45, 48).
[95] See therefore Choi et al. (2010, p. 116); Krauß et al. (2015, p. 49).
[96] See therefore Choi et al. (2010, p. 121); Krauß et al. (2015, p. 49).
[97] See therefore Asthana and Boone (2012, pp. 1, 15); Casterella et al. (2004, p. 123).
[98] See therefore Blankely et al. (2013, p. 18); Eshleman and Guo (2014, p. 135).
[99] See therefore Eshleman and Guo (2014, p. 120).
[100] Overview in Hay (2013, pp. 165-166).
[101] See therefore Seidel (2017, p. 1370).
[102] See therefore Seidel (2017, p. 1370).
[103] See therefore Seidel (2017, p. 1370).

the remediation of an internal control weakness is associated with a decrease in audit fees for that firm.[104]

In the Audit Fee Model, I add several control variables to control for the effect of ΔICW on *LNAFEE*. The control variables resemble companies', auditor, and engagement attributes.[105] Within the companies' attributes, I control for size (*LNTA*, *LNSALES*), complexity (*MERGER*), inherent risk (*INVREC*, *CA*, *INTANG*), profitability (*ROA*, *LOSS*), and leverage and liquidity (*CR*, *LEV*). *LNTA* is the logarithm of the company's total assets, whereas *LNSALES* is the logarithm of sales made by the company. *MERGER* is a dummy variable with the value of 1 if the firm's acquisition value is not equal to zero. *INVREC* is the sum of inventories and receivables divided by total assets. *CA* represents the current-asset ratio and *CR* the current ratio. *INTANG* calculates by dividing intangibles by total assets. *ROA* is the return on assets and calculates by dividing EBIT by total assets. *LEV* is the leverage ratio or long-term debt divided by total assets. LOSS is a dummy variable with the value of 1 if the current year's net income is negative. For *LNTA*, *LNSALES*, *MERGER*, *INVREC*, *CA*, *INTANG*, *LOSS*, and *LEV*, I expect a significant positive association with audit fees. The intuition behind this is that audit fees increase as the scope of audit increases due to a larger and more complex company.[106] For *ROA* and *CR*, I expect a negative correlation with audit fees, as profitable and liquid companies are usually less likely to be affected by internal control weaknesses.[107] Within the auditors' attributes and engagement attributes, I control for auditor quality (*BIG4*), busy season (*BUSY*), and audit problems (*OPINION*). *BIG4* is a dummy variable that takes the value of 1 if the auditor is one of the big four audit firms (i.e., KPMG, EY, Deloitte, PWC). *BUSY* is a dummy variable that takes the value of 1 if the fiscal year ends in December. *OPINION* is a dummy variable with a value of 1 if the auditor issues a going concern audit opinion. For the *BIG4* variable, I expect a positive association since there is empirical evidence that audit fee premiums for Big4 clients exist.[108] The *BUSY* variable is also positively associated because audits are more expensive at the end of the year.[109] In addition, a positive association is expected if the auditor issues a going-concern opinion (*OPINION*), as this is associated with a more significant effort.[110] I also control for fixed effects in different industries (*INDCON*) and different years (*YEAR*).

3.1.2. Restatement Model

Hypothesis 2a addresses the relationship between abnormal audit fees and audit risk. Moreover, hypothesis 2b addresses the relationship between the remediation of internal control weaknesses and audit risk. The problem with examining relationships concerning audit risk is that the assurance auditors provide is not observable.[111] Thus, it is challenging to measure audit risk directly. Therefore, I use the probability of restatements as a proxy for audit risk. The strengths of this proxy are low measurement errors, and it serves as solid evidence for poor audit quality respectively high audit risk.[112] In line with that, Rajgopal et al. (2021) report that restatements are a helpful predictor of poor audit quality.[113] Furthermore, the relationship between restatements and audit risk is positive. In other words, the higher the probability of a restatement, the higher the audit risk. Analogous to the Audit Fee Model, I rely on previous models concerning restatements. In these Restatement Models, the dependent variable is a dummy variable that takes the value of 1 if the company has issued a financial statement restatement for the respective year.[114] The independent explanatory variables are the abnormal audit fees (*ABAFEE*) and the remediation of internal control weaknesses (ΔICW). Consequently, I examine hypotheses 2a and 2b by the following regression model:[115]

Restatement Model:

$$\begin{aligned} REST_{i,t} = {} & \beta_0 + \beta_1 ABAFEE_{i,t} + \beta_2 \Delta ICW_{i,t} + \beta_3 LNTA_{i,t} \\ & + \beta_4 EARNINGSGROW_{i,t} + \beta_5 XSCORE_{i,t} \\ & + \beta_6 LOSS_{i,t} + \beta_7 ROA_{i,t} + \beta_8 MERGER_{i,t} \\ & + \beta_9 FREEC_{i,t} + \beta_{10} LEV_{i,t} + \beta_{11} BIG4_{i,t} \\ & + \beta_{12-20} INDCON_{i,t} + \beta_{21-30} YEAR_{i,t} + \varepsilon_{i,t} \end{aligned}$$

The definition of ΔICW is the same as in the Audit Fee Model. Therefore, I expect a negative association between the remediation of internal control weaknesses and the probability of restatements.[116] The intuition behind this is that due to the remediation of internal control weaknesses, audit risk decreases, and thus the probability of restatements decreases. Abnormal audit fees equal the differences between the actual audit fees paid and the expected audit fees. Thus, the residuals from the Audit Fee Model correspond to the abnormal audit fees (*ABAFEE*). Therefore, hypothesis 2a proposes that abnormal audit fees are negatively associated with audit risk and thus negatively with the probability of restatement.[117]

Analogous to the Audit Fee Model, I add control variables in the Restatement Model to control for the effect of the explanatory variables (ΔICW, *ABAFEE*) on *REST*. The control variables concern the company's size (*LNTA*), complexity (*MERGER*), inherent risk (*XSCORE*, *EARNINGSGROW*,

[104] See therefore Albring et al. (2018, p. 499); Elder et al. (2009, p. 545); Hay (2013, pp. 170-171); Krishnan et al. (2011, p. 1).
[105] See therefore Hay (2013, pp. 170-171).
[106] Analogous to Doyle et al. (2007, p. 193).
[107] See therefore Doyle et al. (2007, p. 193).
[108] See therefore Hoag and Hollingsworth (2011, p. 196).
[109] See therefore Hay (2013, p. 173).
[110] See therefore Hay (2013, p. 171).
[111] See therefore DeFond and Zhang (2014, p. 283); Seidel (2017, p. 1370).
[112] See therefore DeFond and Zhang (2014, p. 285).
[113] See therefore Rajgopal et al. (2021, p. 595).
[114] See therefore Pittman and Zhao (2021, p. 133); Rajgopal et al. (2021, p. 573).
[115] See therefore DeFond and Zhang (2014, p. 290).
[116] Analogous to the findings of Guragai and Hutchinson (2019, p. 363).
[117] See therefore Doogar et al. (2015, p. 1278).

FREEC), profitability (*ROA*, *LOSS*), and leverage (*LEV*) as well as the quality of the auditor (*BIG4*). The control variable *XSCORE* is equivalent to Zmijewski's (1984) financial distress model.[118] Within this model, values greater than zero predict that the company will go bankrupt. Therefore, I expect a positive association between *XSCORE* and *REST*.[119] *FREEC* captures the demand for external financing. It measures the company's ability to cover its capital expenditures through holding assets by dividing net cash flow from operating activities less capital expenditures by total assets.[120] As previous findings show, I expect a positive relationship between *FREEC* and *REST*.[121] Finally, *EARNINGSGROW* is a dummy variable that takes the value 1 if the net income in $t = 1$ is greater than in $t = 0$. This control variable addresses the pressure managers face to retain earnings growth.[122] Furthermore, earnings pressure leads managers to opportunistically accounting procedures that eventually lead to restatements.[123] Therefore, I expect a positive association between *EARNINGSGROW* and *REST*. Like in the Audit Fee Model, additionally, I control for fixed effects in different industries (*INDCON*) and different years (*YEAR*).

3.1.3. Sample and Data Selection

My analyses use data from several 10-K, 10-Q, and 8-Q reports and proxy statements between 2010 to 2020, with overall 66,500 observations from 9,869 individual companies. To create the sample, I first eliminated all firms whose restatement announcement was not later than 2018, leaving 65,677 observations. This filtering is standard in the literature, as previous studies find an average lag of two years between the end of the reporting period and the restatement announcement.[124] In the following, I eliminate firms that do not have an internal report and have a different accounting standard than the United States Generally Accepted Accounting Principles (U.S. GAAP), leaving 25,300 observations. Next, I eliminate observations with no audit fee data or missing data to calculate regression model variables, leaving 17,061 observations. In the final step, I eliminate financial firms and regulated firms. That is because financial firms have different financial structures and, therefore, different audit production functions.[125] At the same time, the capital regulation is different for regulated firms, which also separates them from other industries.[126] Panel A from Table 1 shows that this procedure leaves 15,326 observations from 2,609 individual firms.

I identify all observations that show remediation of internal control weaknesses from this filtered data set, resulting in 225 observations representing the treatment group for my analyses. Then, to form an appropriate control group, I randomly pick 225 observations that show internal control weaknesses but no remediation. Finally, Panel B of Table 1 shows that my final sample has 450 observations.

The sample indicates that most firms are affected for two years by restatements (see Panel C of Table 1). This indication aligns with my condition to leave a lag of two years between the end of the reporting period and the restatement announcement. Furthermore, from the 450 sample observations, 102 exhibit a restatement. This ratio of approximately 22.7% is appropriate, as restatements are rare events.[127]

The descriptive analysis in Table 2 gives additional information about the sample. The average company has total assets worth approximately $564.03 million and makes sales of $439.66 million. Regarding the distribution of the different asset types, approximately half of the assets correspond to current assets for the average company. In terms of leverage and liquidity, the current ratio of approximately 2.52 shows that the average company has the financial resources to remain solvent in the short term. Additionally, a long-term debt-to-assets ratio (*LEV*) of only approximately 23.9% shows that, on average, the companies are in a healthy position. However, in nearly 45% of the cases, the company has incurred a loss. The mean return on assets is negative, implying that the average company is not profitable. Moreover, the average company has an increased inherent risk. A negative mean of the *FREEC* variable indicates that the average company cannot cover its capital expenditures through holding assets. In contrast, the negative mean of *XSCORE* from the Zmijewski (1984) model means that the average company is predicted not to go bankrupt.[128] Lastly, the earnings of the average company grow with a probability of approximately 55.6%.

Regarding auditors, on average, companies pay approximately $1.42 million for audit fees. Furthermore, in approximately 65% percent of the cases, the auditor is a Big4 auditor.

In addition, Table 3 gives a summary of the correlations among regression. These indicate possible associations for the Audit Fee and Restatement Model.

Table 3 states a correlation coefficient of -0.04 between *LNAFEE* and ΔICW, indicating a negative relationship. This association would be in line with the first hypothesis. Furthermore, a correlation of -0.06 between *REST* and *ABAFEE* also suggests a negative relationship in line with hypothesis 2a. Finally, the correlation coefficient of -0.12 between *REST* and ΔICW indicates the relationship expected in hypothesis 2b. In summary, the correlations among the regression variables are consistent with the expected relationships.

[118] See therefore Zmijewski (1984, pp. 65-66).
[119] See therefore Viciwati (2020, p. 797).
[120] See therefore Aier et al. (2005, p. 129); Blankely et al. (2012, pp. 84-85).
[121] See findings of Aier et al. (2005, p. 133); Blankely et al. (2012, p. 92).
[122] See therefore Blankely et al. (2012, p. 85).
[123] See therefore Garcia Osma (2008, p. 117).
[124] See therefore Blankely et al. (2012, p. 85); Kester et al. (2013, pp. 163-164); Lobo and Zhao (2013, pp. 1394-1395); Pittman and Zhao (2021, p. 134).
[125] See therefore Albring et al. (2018, p. 492); Doogar et al. (2015, p. 1260); Fields et al. (2004, p. 53); Stein et al. (1994, p. 128).
[126] See therefore Albring et al. (2018, p. 492).
[127] See therefore DeFond and Zhang (2014, p. 284).
[128] See therefore Viciwati (2020, p. 797); Zmijewski (1984, pp. 65-66).

Table 1: Sample Attrition and Composition

Sample Attrition and Composition

Panel A: Sample Attrition

	Firms	**Observations**
Restatement Period 2010-2018	9,869	65,677
Less:		
No Internal Report	(5,000)	(36,036)
Different Accounting Standard	(715)	(4,341)
No Audit Fee Data	(0)	(83)
Missing Calculation Data	(1,237)	(8,156)
Financial and Regulated Firms	(308)	(1,735)
	2,609	15,326

Panel B: Sample Composition

Change in Internal Control Quality (= Treatment Group)	225
No Change in Internal Control Quality (= Control Group)	225
	450

Panel C: Total Years Affected by Restatements

1 year	11
2 years	25
3 years	22
4 years	13
5 years	18
6 years	10
7 years	2
9 years	1
Total Number of unique Restatements	102

3.2. Results and Interpretation

Analogous to Blankely et al. (2012), the first regression of this analysis examines the relationship between the remediation of internal control weaknesses and audit fees and forms the residuals representing the abnormal audit fees from the audit fee model.[129] Table 4 shows the regression results.

The model shows moderate significance ($R^2 = 69.14\%$) by 450 observations. For all years except 2010, 2015, 2016, and 2020 the year controls show a significant negative coefficient (not tabulated). I only find a significant negative coefficient for the food and transportation industry regarding fixed industry effects. All other industries also show a negative coefficient but no significance. The median of the residuals is positive, proposing that the amount of audit fees companies pay is slightly higher than the amount of audit fees the regression model is expecting.

Furthermore, not all control variables are significant. The regression variables *BUSY*, *LEV*, *ROA*, *INVREC*, *MERGER*, and *LNSALES* do not indicate an effect on audit fees. Contrary to expectations, *OPINION* is significantly negative.[130] This result suggests that audit fees decrease if the auditor reports

[129] See therefore Blankely et al. (2012, p. 87).

[130] See therefore Hay (2013, p. 171).

Table 2: Descriptive Analysis

Descriptive Analysis

Cont. Variables	Mean	Std. Dev.	Min	Q1	Median	Q3	Max
LNAFEE	0.3513	1.1925	-4.6861	-0.4125	0.3837	1.0462	4.1658
ABAFEE	0.0000	0.6625	-3.9301	-0.2791	0.0427	0.3888	2.4587
LNTA	6.3351	1.5916	0.9314	5.3600	6.3973	7.3711	11.5273
LNSALES	6.0860	1.7856	-1.7960	5.0800	6.2500	7.2370	10.4410
CR	2.5220	2.2897	0.1550	1.2700	1.8660	2.9300	24.8450
CA	0.4729	0.2219	0.0263	0.2893	0.4563	0.6313	0.9791
INVREC	0.2543	0.1573	0.0000	0.1371	0.2340	0.3612	0.7896
ROA	-0.0262	0.3368	-4.7344	-0.0193	0.0467	0.0841	0.7384
LEV	0.2392	0.2456	0.0000	0.0148	0.1874	0.3796	1.4931
INTANG	0.2434	0.2308	0.0000	0.0365	0.1801	0.4073	0.8489
FREEC	-0.0117	0.2883	-1.4606	-0.0466	0.0192	0.0599	4.7384
XSCORE	-0.5776	3.6286	-18.1406	-2.2538	-0.5776	0.1564	37.5157

Indicator Variables	Mean	Std. Dev.	0	1
ΔICW	0.5000	0.5006	225	225
REST	0.2267	0.4191	348	102
EARNINGSGROW	0.5556	0.4975	200	250
LOSS	0.4467	0.4977	249	201
MERGER	0.3489	0.4771	293	157
BUSY	0.6933	0.4616	138	312
OPINION	0.7133	0.4527	129	321
BIG4	0.6511	0.4771	157	293

a going-concern opinion. In line with previous studies, like DeFond and Zhang (2014), auditor quality and size (BIG4) is strong and significant positive with audit fees.[131] As DeFond and Zhang (2014) state, this result indicates that companies pay a premium on the fee if their auditor company is either KPMG, Ernst & Young, PricewaterhouseCoopers, or Deloitte. The current ratio (*CR*) is significantly negatively associated with audit fees.[132] Thus, the better the company can pay its short-term liabilities, the lower the audit fees, suggesting that auditors pay less attention to highly liquid firms. In terms of profitability, a negative net income (*LOSS*) increases audit fees by approximately $1.28 million (exponential value of 0.250). The intangibles ratio and the current-assets ratio are also significantly positively associated with audit fees, proposing that the ratio of specific asset types to total assets influences audit fees. At last, company size in the form of total assets is strongly significant and positively associated with audit fees. This result aligns with the consensus in the literature that audit fees increase as the work of scope that comes with the size of the company increases.[133]

In line with the analyses of Hammersley et al. (2012), Hoag and Hollingsworth (2011), Mitra (2009), and Munsif et al. (2011), I find a significant negative association between the remediation of internal control weaknesses and audit fees.[134] This result is consistent with hypothesis 1 and indicates that if the company remediates its internal control weakness, the audit fees decrease by approximately $1.19 million (exponential value of 0.175) in the year of remediation. Furthermore, the result purposes that an investment in the remediation of internal control weaknesses decreases control risk. The reason for that is that according to the Audit

[131] See therefore DeFond and Zhang (2014, p. 300).
[132] See therefore Hay (2013, p. 170).
[133] See therefore Albring et al. (2012, p. 388); Bae et al. (2021, p. 606); Calderon et al. (2012, pp. 696-697); Hay (2013, p. 170); Hoag and Hollingsworth (2011, p. 191); Krishnan et al. (2011, p. 14); Mitra (2009, p. 379).
[134] See therefore Hammersley et al. (2012, p. 73); Hoag and Hollingsworth (2011, p. 173); Mitra (2009, p. 369); Munsif et al. (2011, p. 87).

Table 3: Correlations among Regression Variables

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1.LNAFEE	1.00																			
2.ABAFEE	0.56	1.00																		
3.LNTA	0.75	0.00	1.00																	
4.LNSALES	0.70	0.00	0.89	1.00																
5.EARNINGSGROW	-0.02	0.03	0.01	0.05	1.00															
6.CR	-0.23	0.00	-0.18	-0.26	0.00	1.00														
7.CA	-0.21	0.00	-0.38	-0.29	0.02	0.39	1.00													
8.INVREC	0.03	0.00	-0.07	0.19	0.03	-0.02	0.51	1.00												
9.ROA	0.24	0.00	0.41	0.49	0.06	0.03	-0.10	0.14	1.00											
10.LEV	0.34	0.00	0.40	0.34	-0.02	-0.17	-0.39	-0.17	0.05	1.00										
11.INTANG	0.27	0.00	0.28	0.19	-0.04	-0.18	-0.56	-0.30	0.11	0.18	1.00									
12.FREEC	0.08	-0.01	0.18	0.19	0.09	0.01	-0.04	0.11	0.36	-0.03	0.07	1.00								
13.XSCORE	0.03	0.04	-0.12	-0.09	-0.02	-0.29	0.00	-0.04	-0.59	0.39	-0.06	-0.43	1.00							
14.ΔICW	-0.04	0.00	-0.09	0.07	0.03	0.12	0.04	-0.02	0.01	-0.01	-0.02	0.07	-0.10	1.00						
15.REST	0.02	-0.06	-0.01	0.10	-0.10	-0.05	-0.09	0.04	0.06	0.06	0.03	0.00	-0.01	-0.12	1.00					
16.LOSS	-0.04	0.00	-0.31	-0.30	-0.28	-0.06	-0.03	-0.12	-0.36	0.08	0.07	-0.27	0.29	-0.11	0.00	1.00				
17.MERGER	0.21	0.00	0.28	0.19	-0.07	-0.11	-0.24	-0.12	0.13	0.06	0.33	0.07	-0.06	-0.04	0.06	-0.07	1.00			
18.BUSY	0.05	0.00	0.01	-0.05	-0.03	-0.14	-0.12	-0.12	-0.04	0.08	0.07	-0.04	0.08	0.01	-0.04	0.01	0.01	1.00		
19.OPINION	-0.15	0.00	-0.07	-0.02	0.01	0.06	0.02	0.03	0.13	-0.07	-0.04	0.04	-0.10	0.03	0.03	-0.09	0.02	-0.04	1.00	
20.BIG4	0.55	0.00	0.51	0.49	-0.05	-0.14	-0.09	0.05	0.15	0.25	0.10	0.02	0.06	0.12	0.04	-0.13	0.06	0.03	-0.10	1.00

Table 4: Results from the Audit Fee Model

Results: Audit Fee Model

Residuals:

	Min	Q1	Median	Q3	Max
	-3.930	-0.279	0.043	0.389	2.459

Coefficients:

	Estimate	Std. Error	t-Value	p-Value	
Intercept	-2.930	0.579	-5.063	0.000	***
$\triangle$ICW	**-0.175**	**0.069**	**-2.533**	**0.012**	**
LNTA	0.430	0.062	6.924	0.000	***
LNSALES	0.010	0.060	1.634	0.103	
MERGER	0.082	0.075	1.094	0.275	
INVREC	0.290	0.332	0.872	0.384	
CA	0.676	0.256	2.639	0.009	***
INTANG	0.551	0.198	2.786	0.006	***
ROA	-0.196	0.124	-1.585	0.114	
LOSS	0.250	0.074	3.359	0.001	***
CR	-0.036	0.018	-1.957	0.051	*
LEV	0.168	0.016	1.040	0.299	
BIG4	0.447	0.084	5.342	0.000	***
BUSY	0.117	0.079	1.487	0.138	
OPINION	-0.174	0.083	-2.082	0.038	**
Fixed Industry Effects	Included				
Fixed Year Effects	Included				
Multiple R^2	0.6914				
Observations	450				

Notes: ***, **, and * indicate significance level at 1, 5, and 10 percent.
Bold text indicates variable of interest.

Risk Model, lower audit fees relate to higher detection risk, which is just appropriate if the control risk of the company is low.[135] In summary, I conclude that the remediation of internal control weaknesses is an effective way to improve the internal control systems and, thus, an effective way to decrease control risk. Moreover, companies must pay a lower amount of audit fees if they remediate their internal control weaknesses.

Analogous to Blankely et al. (2012), the second regression of my analysis examines the relationship between the probability of restatements and abnormal audit fees on the one hand and the remediation of internal control weaknesses on the other hand.[136] Table 5 shows the regression results.

The model shows a multiple R^2 of 13.78% by 450 observations. This significance is in line with other Restatement Models of previous studies.[137] For all years except 2010, 2019, and 2020 the year controls show a significant positive coefficient (not tabulated). The industry controls do not show a significant association with the occurrence of restatements (not tabulated).

Only *EARNINGSGROW* exhibits a significant relationship with the dependent variable *REST* in terms of control variables. In more detail, if the company's net income has increased from last year to the current year, then the probability of restatements decreases by approximately 7.7 percentage points. This result suggests that successful companies are

[135] See therefore Elder et al. (2009, p. 545); PCAOB (2010, p. 64).
[136] See therefore Blankely et al. (2012, p. 84).
[137] See therefore Blankely et al. (2012, p. 92); Krauß et al. (2015, p. 76); Lawrence et al. (2018, p. 151).

Table 5: Results from the Restatement Model

Results: Restatement Model

Residuals:

	Min	Q1	Median	Q3	Max
	-0.612	-0.262	-0.164	-0.090	0.968

Coefficients:

	Estimate	Std. Error	t-Value	p-Value	
Intercept	-0.384	0.327	-1.173	0.241	
ABAFEE	**-0.037**	**0.029**	**-1.301**	**0.194**	
∆ICW	**-0.095**	**0.040**	**-2.362**	**0.019**	**
LNTA	0.021	0.017	1.239	0.216	
EARNINGSGROW	-0.077	0.042	-1.860	0.064	*
XSCORE	-0.005	0.008	-0.571	0.568	
LOSS	-0.000	0.045	-0.008	0.994	
ROA	0.044	0.083	0.525	0.560	
MERGER	0.019	0.042	0.440	0.660	
FREEC	-0.019	0.076	-0.250	0.802	
LEV	0.080	0.105	0.767	0.444	
BIG4	0.006	0.049	0.122	0.903	
Fixed Industry Effects	Included				
Fixed Year Effects	Included				
Multiple R^2	0.1378				
Observations	450				

Notes: ***, **, and * indicate significance level at 1, 5, and 10 percent.
Bold text indicates variable of interest.

less likely to announce and disclose a restatement of financial statements. In other words, successful companies have a lower audit risk. Contrary to expectations, *LNTA*, *XSCORE*, *LOSS*, *ROA*, *MERGER*, *FREEC*, *LEV*, and *BIG4* do not show significance.

Not in line with hypothesis 2a, there is no significant association between abnormal audit fees (*ABAFEE*) and the likelihood of restatements (*REST*), respectively, audit risk and audit quality. Similar to the sign of their correlation in Table 3, the sign of the regression coefficient is negative. The negative sign would indicate that abnormal audit fees are associated with a lower probability of restatements, lower audit risk, and thus higher audit quality, but since there is no significance, abnormal audit fees have no impact on audit risk and audit quality. This result contradicts the results of studies concerning the effort view and the economic bonding view.[138] Thus, abnormal audit fees represent no additional audit effort and no economic bond to the auditor, making the auditor less independent. In the overall perspective, this would suggest for companies that changes in audit fees due to changes in internal control systems have no impact on audit risk.

In contrast to this proposal, I find a significant negative association between the remediation of internal control weaknesses (∆*ICW*) and the probability of restatements (*REST*). This result suggests that the remediation of internal controls is associated with a lower probability of restatements. More precisely, the likelihood of restatement decreases by 9.5 percentage points if the company remediates its internal control weaknesses. Further, the remediation of internal control weaknesses reduces audit risk, consistent with hypothesis 2b. From a general point of view, this implies that the remediation of internal control weaknesses effectively reduces audit risk and improves audit quality. Furthermore, assuming a close relationship between audit quality and financial reporting quality, this result implies that a company can significantly improve its financial reporting

[138] See therefore Blankley et al. (2014, p. 30); Eshleman and Guo (2014, p. 120).

quality by remediating its internal control weaknesses.[139] Consequently, the remediation of internal control weaknesses is associated with improving the relationship between the company and its investors. Due to the remediation, financial statements are more credible and thus improve investor confidence.[140]

In conclusion, the remediation of internal control weaknesses is associated with a decrease in audit fees and audit risk and thus an improvement of audit and financial reporting quality. These effects of changes in internal control systems on audit risk represent important implications for companies, as internal control weaknesses and restatements, in particular, impact the company and its value.[141]

3.3. Additional Analysis

Previous literature suggests that audit fees consist of an effort and a risk premium component.[142] As a result of this risk premium component, the interpretation of changes in audit fees is ambiguous.[143] More specifically, it is unknown how much audit fees change due to changes in effort or risk premium. Therefore, to interpret the Audit Fee Model appropriately, I conduct an additional analysis examining risk premium effects. Analogous to Albring et al. (2012), I use an adjusted Audit Fee Model.[144] The explanatory variables in this model are *REMY1*, *REMY2*, and *REMY3*. These variables are dummy variables that take the value of 1 if the firm remediated an internal control weakness and it is the first, second, or third year after the disclosure of the internal control weakness.[145] Moreover, this definition assumes that the remediation of the internal control weaknesses is in the year following their disclosure. Therefore, *REMY1* takes the value of 1 if it is the year the firm remediated its internal control weakness. Different from the previous models in this analysis is the sample. The treatment group consists of the remediation and the following two years, resulting in 108 observations. The control group represents firms that never disclosed an internal control weakness.[146] Overall, the sample has 216 observations. Thus, the regression equation for this additional analysis is:

Audit Risk Premium Analysis:

$$\begin{aligned} LNAFEE_{i,t} = {} & \beta_0 + \beta_1 REMY1_{i,t} + \beta_2 REMY2_{i,t} \\ & + \beta_3 REMY3_{i,t} + \beta_4 LNTA_{i,t} + \beta_5 LNSALES_{i,t} \\ & + \beta_6 MERGER_{i,t} + \beta_7 INVREC_{i,t} + \beta_8 CA_{i,t} \\ & + \beta_9 INTANG_{i,t} + \beta_{10} ROA_{i,t} + \beta_{11} LOSS_{i,t} \\ & + \beta_{12} CR_{i,t} + \beta_{13} LEV_{i,t} + \beta_{14} BIG4 \\ & + \beta_{15} BUSY_{i,t} + \beta_{16} OPINON_{i,t} \\ & + \beta_{27-25} INDCON_{i,t} + \beta_{26-35} YEAR_{i,t} + \varepsilon_{i,t} \end{aligned}$$

The results in Table 6 are partially consistent with the findings of Albring et al. (2012).[147] *REMY1* is significantly positive, indicating that the audit fees remain high in the year of remediation. Regarding the results of ΔICW in the Audit Fee Model, this result indicates that audit fees decrease due to the remediation of internal control weaknesses but do not decrease to the level that the firm would have had in the absence of internal control weaknesses. Moreover, this implies a risk premium component in the audit fees in the year of remediation. *REMY2* is also positively significant and thus suggesting that audit fees remain high one year after remediation. Furthermore, this implies that the risk premium components are also present in the second year after internal control disclosure. The coefficient of *REMY2* is smaller than that of *REMY1*. This finding indicates that the risk premium decreases in the year after the remediation. There is no risk premium effect in the second year following the remediation of internal control weaknesses since *REMY3* is insignificant. These results conclude that the results of the Audit Fee Model contain a risk premium component, and therefore the interpretation of *LNAFEE* is ambiguous.[148]

3.4. Robustness Tests

The results indicate that the remediation of internal control weaknesses is associated with an increase in audit fees and audit quality. For the underlying regression models, I use a sample of 450 observations. I filtered the data set only for essential parts to obtain this sample, such as observations with the internal control report. This procedure keeps the number of observations as high as possible to maintain a reasonable statistical power. In contrast, Blankely et al. (2012) filter the data set even more to exclude additional effects on audit fees and restatements.[149] In more detail, they eliminate firms with multiple restatements and non-Big4 auditors. With these executions, they want to guarantee a certain homogeneity of audit quality in the sample.[150] Therefore, I test my results on robustness concerning this additional sample filtering. Consequently, the treatment and control group size decreases to 59 observations each, resulting in overall 118

[139] See therefore DeFond and Zhang (2014, p. 287).
[140] See therefore Hammersley et al. (2012, p. 74).
[141] See therefore Pittman and Zhao (2021, p. 129).
[142] See therefore Albring et al. (2012, p. 377); DeFond and Zhang (2014, p. 290); Munsif et al. (2011, p. 89); Simunic (1980, p. 169).
[143] See therefore DeFond and Zhang (2014, p. 290).
[144] See therefore Albring et al. (2012, p. 382).
[145] See therefore Albring et al. (2012, p. 383).
[146] See therefore Albring et al. (2012, p. 377).
[147] See therefore Albring et al. (2012, p. 389).
[148] In line with DeFond and Zhang (2014, p. 290).
[149] See therefore Blankely et al. (2012, p. 85).
[150] See therefore Blankely et al. (2014, p. 37); Blankely et al. (2012, p. 85).

Table 6: Audit Fee Premium Analysis

Results: Audit Fee Premium Analysis

Residuals:

	Min	Q1	Median	Q3	Max
	-2.854	-0.217	0.036	0.294	0.095

Coefficients:

	Estimate	Std. Error	t-Value	p-Value	
Intercept	-3.483	0.521	-6.687	0.000	***
REMY1	**0.265**	**0.113**	**2.351**	**0.020**	**
REMY2	**0.201**	**0.116**	**1.723**	**0.087**	*
REMY3	**0.046**	**0.120**	**0.379**	**0.705**	
LNTA	0.376	0.093	4.060	0.000	***
LNSALES	0.066	0.098	0.675	0.500	
MERGER	0.134	0.094	1.429	0.155	
INVREC	-0.114	0.391	-0.293	0.770	
CA	0.694	0.381	1.820	0.070	*
INTANG	0.190	0.283	0.672	0.502	
ROA	0.071	0.330	0.215	0.830	
LOSS	0.203	0.100	2.024	0.044	**
CR	-0.071	0.021	-3.358	0.000	***
LEV	0.383	0.180	2.130	0.034	**
BIG4	0.635	0.107	5.954	0.000	***
BUSY	-0.032	0.089	-0.364	0.716	
OPINION	-0.004	0.109	-0.038	0.970	
Fixed Industry Effects	Included				
Fixed Year Effects	Included				
Multiple R^2	0.8122				
Observations	216				

Notes: ***, **, and * indicate significance level at 1, 5, and 10 percent.
Bold text indicates variable of interest.

sample observations. The conduct of the two regression models is the same as in my previous analysis. The results of the explanatory variables are in Table 7.

The result of the adjusted Audit Fee Model confirms the previous results. Therefore, it suggests a robust negative association between the remediation of internal control weaknesses and audit fees. The results of the adjusted Restatement Model deviate from the previous results. On the hand, I now find a significant negative association between abnormal audit fees and restatements which confirms the negative sign of the correlation coefficient in my first analysis. On the other hand, I find deviating results for ΔICW. One reason for this could be that the standard errors are larger than the coefficient itself, resulting from a too small sample. Thus, this robustness test shows that the applied procedure of filtering the data set to maintain a reasonable statistical power is appropriate.

In addition, Blankely et al. (2012) only consider restatements with a negative effect.[151] In this analysis sample, there are 83 negative but also 19 positive effects of restatements. Therefore, I check if the results from the Restatement Model change by only considering adverse restatements. Following that, the size of the treatment and control group decreases to 212 observations each, resulting in 424 observations over-

[151] See therefore Blankely et al. (2012, p. 86).

Table 7: Results of Robustness Test on Sample

Results: Audit Fee Model					
Coefficients:					
	Estimate	Std. Error	t-Value	p-Value	
ΔICW	**0.186**	**0.104**	**-1.785**	**0.078**	*
Control Variables	Included				
Fixed Industry Effects	Included				
Fixed Year Effects	Included				
Multiple R^2	0.8226				
Observations	118				
Results: Restatement Model					
Coefficients:					
	Estimate	Std. Error	t-Value	p-Value	
ABAFEE	**-0.167**	**0.061**	**-2.723**	**0.008**	***
ΔICW	**0.032**	**0.056**	**0.580**	**0.563**	
Control Variables	Included				
Fixed Industry Effects	Included				
Fixed Year Effects	Included				
Multiple R^2	0.2635				
Observations	118				

Notes: ***, **, and * indicate significance level at 1, 5, and 10 percent.
Bold text indicates variable of interest.

all.[152] Table 8 shows the results.

The results of this robustness test are consistent with the results of the Restatement Model in the analysis. This result indicates that it is not mandatory only to consider negative restatements when regarding the associations of abnormal audit fees and the remediation of internal control weaknesses on the probability of restatements.

Furthermore, I follow the suggestion of DeFond and Zhang (2014) and examine audit quality with a different proxy.[153] Therefore, I use discretionary accruals to examine the relations between audit quality, audit fees, and the remediation of internal control weaknesses. Thus, analogous to Eshleman and Guo (2014), I calculate total accruals using the Jones (1991) Model, as modified by Ball and Shivakumar (2006), to finally estimate the residuals from this model corresponding to discretionary accruals.[154]

Total Accruals:

$$TACC_{i,t} = \beta_0 + \beta_1 \Delta SALE_{i,t} + \beta_2 PPE_{i,t} + \beta_3 OCF_{i,t} + \beta_4 DCF_{i,t} + \beta_5 OCF_{i,t} * DCF_{i,t}$$

I then replaced the restatement model's dependent variable *REST* with the new proxy for audit quality *DAC* (discretionary accruals). Table 8 reports the results of this procedure.

The results from the Discretionary Accruals Model are contrary to the Restatement Model. Nevertheless, I adhere to the results of the Restatement Model since discretionary accruals suffer from significant measurement errors, which ultimately lead to biases.[155] These measurement errors seem to be particularly the case when the sample firms are not homogeneous within an industry.[156] However, in the context of

[152] The reduction in observations is greater than 19 to maintain that the treatment and control group are evenly sized. Within the control group, there are 13 positive restatement observations whereas 6 observations in the treatment group. 225 control group observations less 13 equal 212 observations each.

[153] See therefore DeFond and Zhang (2014, p. 276).

[154] See therefore Ball and Shivakumar (2006, p. 226); Eshleman and Guo (2014, p. 126); Jones (1991, p. 221).

[155] See therefore DeFond and Zhang (2014, p. 288).

[156] See therefore Dopuch et al. (2012, p. 408).

Table 8: Negative Restatements in the Restatement Model

Results: Restatement Model					
Coefficients:					
	Estimate	Std. Error	t-Value	p-Value	
ABAFEE	**-0.016**	**0.031**	**-0.516**	**0.606**	
ΔICW	**-0.085**	**0.039**	**-2.189**	**0.029**	**
Control Variables	Included				
Fixed Industry Effects	Included				
Fixed Year Effects	Included				
Multiple R^2	0.1241				
Observations	424				

Notes: ***, **, and * indicate significance level at 1, 5, and 10 percent.
Bold text indicates variable of interest.

this study, I cannot rule out the possibility that the industries of the sample firms may or may not be heterogeneous.

4. Discussion of Results

In this analysis, I examine the effects of changes in internal control systems on audit risk. In this study, using the Audit Fee Model, I find that the remediation of internal control weaknesses associates with a decrease in audit fees. Moreover, using the Restatement Model, I find no significant association between abnormal audit fees and the probability of restatements. However, I report a significant negative relationship between the remediation of internal control weaknesses and restatements. This result indicates that the remediation decreases audit risk and thus increases audit quality. At this point, a few limitations, and open questions, are worth mentioning.

At first, limitations concerning the sample. Instead of picking the observations of the control group randomly, there would also be the possibility of conducting a propensity score matching process. The studies from Guragai and Hutchinson (2019) and Mitra et al. (2017) also examine associations concerning internal control weaknesses and use a propensity score matching process to determine a matched-pair sample of the treatment group.[157] The advantage of this approach is that the treatment group is better represented, and thus the effect that comes from the treatment group can be better shown. However, with my data set, it is very unclear if this approach would improve the representation of the control group. This concern arises because the propensity score is obtained from a regression with the dependent variable resembling my Δ*ICW* variable would be low. This assertion turns out from the fact that the correlations of all other variables with Δ*ICW* are weak (Table 3). Another limitation concerning the sample is the general amount of remediations of internal control weaknesses in the data set. Before filtering the data set, only approximately 2.2%, and after filtering, approximately 2.8% of the observations show remediation of internal control weaknesses (not tabulated). Finally, this small number of observations eventually leads to high standard errors of the variables in the descriptive analysis (Table 2), which affects their statistical significance.[158]

Second, limitations concerning the Audit Fee Model. Due to data set constraints, it was impossible to determine several variables for the Audit Fee Model. Analyses concerning Audit Fee Models often include variables like the number of foreign subsidiaries, foreign assets, and the number of business segments to depict a company's complexity.[159] Moreover, corporate governance variables like the number of outside directors are not in the data set. The use of these additional variables could increase the representation of the model. Another limitation associated with the audit fee model is the interpretation of audit fees. As shown in the additional analysis part of this study, audit fees consist of an effort and a risk premium component.[160] Moreover, audit fees still include a risk premium even after the remediation of internal control weaknesses. The risk premium in audit fees after remediation limits the Δ*ICW*'s interpretation in the way that the $1.19 million is not the difference of audit fees between the case of the company disclosing and the case of never disclosing an internal control weakness. Thus, I can only interpret the regression coefficient of Δ*ICW* with caution. Another general limitation is the high standard error in the Audit Fee Model,

[157] See therefore Guragai and Hutchinson (2019, p. 366); Mitra et al. (2017, p. 246).

[158] Analogous to DeFond and Zhang (2014, p. 284).

[159] See therefore Blankely et al. (2012, p. 83); DeFond and Zhang (2014, p. 291); Hay (2013, p. 170).

[160] See therefore Simunic (1980, p. 169).

Table 9: Discretionary Accruals Robustness Test

Results: Total Accruals Model

Coefficients:

	Estimate	Std. Error	t-Value	p-Value	
Intercept	-0.061	0.016	-3.751	0.000	***
ΔSALE	0.053	0.039	1.375	0.170	
PPE	-0.075	0.018	-4.100	0.000	***
OCF	0.162	0.072	2.237	0.026	**
DCF	0.109	0.027	4.104	0.000	***
OCF:DCF	0.148	0.087	1.708	0.088	*
Multiple R^2	0.1321				
Observations	450				

Results: Discretionary Accruals Model

Coefficients:

	Estimate	Std. Error	t-Value	p-Value	
ABAFEE	**0.029**	**0.013**	**2.219**	**0.027**	**
ΔICW	**0.025**	**0.018**	**1.352**	**0.177**	
Control Variables	Included				
Fixed Industry Effects	Included				
Fixed Year Effects	Included				
Multiple R^2	0.3260				
Observations	450				

Notes: ***, **, and * indicate significance level at 1, 5, and 10 percent.
Bold text indicates variable of interest.

shown in the results of Table 4.

Third and finally, there are limitations concerning the Restatement Model. Even though restatements as a proxy of audit risk have relatively low measurement errors, they still suffer from interpretation problems. Therefore, the absence of restatements does not directly equate to high audit quality. The reason is that low-quality audits can prevent serious errors but not necessarily detect minor errors, which is necessary for high audit quality.[161] Therefore, regarding my analysis, I can only suggest that the remediation of internal control weaknesses is associated with an increase in audit quality but not with high audit quality. In addition, the reduced likelihood of restatements can only imply a reduction in audit risk since actual audit risk is unobservable.[162] Other limitations come with the explanatory variable *ABAFEE*. On the one hand, *ABAFEE* is dependent on this specific Audit Fee Model in this study.[163] Therefore, this leaves the possibility that different Audit Fee Models with different variables have different residuals and thus a different *ABAFEE*. This concern limits the generalizability of the results in the Restatement Model. On the other hand, there is concern that the residuals from the Audit Fee Model contain an additional error component which would bias the results of the Restatement Model.[164]

5. Conclusion

This study examines the effect of changes in internal control systems on audit risk. This analysis considers the relationships between the remediation of internal control weak-

[161] See therefore DeFond and Zhang (2014, p. 284).
[162] See therefore Seidel (2017, p. 1370).
[163] See therefore Eshleman and Guo (2014, p. 134).
[164] See therefore Eshleman and Guo (2014, p. 134).

nesses, audit fees, and the probability of restatements. Understanding these relations is essential since internal control weaknesses and restatements can impact the firm, its environment, and its value.[165] Significantly, the negative economic consequences are primary drivers for companies to pay attention to internal control systems and audit risk.[166] Therefore, I examine how changes in internal control systems due to remediation of internal control weaknesses affect audit risk. I conduct an audit fee and a restatement regression model to examine this relationship. The Audit Fee Model considers the associations between the audit fees and the remediation of internal controls weaknesses. Whereas the Restatement Model considers the effects of abnormal audit fees and the remediation of internal control weaknesses on the probability of restatements as an audit risk proxy. For both regression models, I use a sample of 450 observations. The sample is composed of a treatment group and a control group. The treatment group represents all observations that show remediation of internal control weaknesses. At the same time, the control group comprises companies that have disclosed an internal control weakness but have not remediated it.

In the Audit Fee Model, I find a significant negative association between audit fees and the remediation of internal control weaknesses. This result indicates that audit fees decrease as the company remediates its internal control weaknesses. The Restatement Model shows no significant relationship between abnormal audit fees and restatements. This result suggests that changes in audit fees due to changes in internal control systems have no impact on audit risk. Finally, the Restatement Model shows a significant negative relationship between the remediation of internal control weaknesses and the probability of restatements. In other words, this means that the remediation of internal controls is associated with a lower probability of restatements and, therefore, with reductions in audit risk and increases in audit quality.

These results are significant as companies can directly influence their internal control systems. Thus, for further research, it might be interesting to examine the direct effect of the remediation of internal control weaknesses on economic variables such as the cost of capital.

References

Aier, J. K., Comprix, J., Gunlock, M. T., & Lee, D. (2005). The Financial Expertise of CFOs and Accounting Restatements. *Accounting Horizon*, *19*(3), 123–135.

Albring, S. M., Elder, R., & Xu, X. (2018). Unexpected Fees and the Prediction of Material Weaknesses in Internal Control Over Financial Reporting. *Journal of Accounting, Auditing & Finance*, *33*(4), 485–505.

Albring, S. M., Elder, R. J., & Keane, M. J. (2012). The effect of the type and number of internal control weaknesses and their remediation on audit fees. *Review of Accounting and Finance*, *11*, 377–399.

Alhadab, M. (2018). Abnormal audit fees and accrual and real earnings management: evidence from UK. *Journal of Financial Reporting and Accounting*, *16*(3), 395–416.

Archambeault, D. S., DeZoort, T. F., & Hermanson, D. R. (2008). Audit Committee Incentive Compensation and Accounting Restatements. *Contemporary Accounting Research*, *24*(4), 965–992.

Asare, S. K., Fitzgerald, B. C., Graham, L., Joe, J. R., Negangard, E. M., & Wolfe, C. J. (2013). Auditors' Internal Control Over Financial Reporting Decisions: Analysis, Synthesis, and Research Directions. *Auditing: A Journal of Practice & Theory*, *32*, 131–166.

Ashbaugh-Skaife, H., Collins, D. W., Kinney Jr., W. R., & LaFond, R. (2008). The Effect of SOX Internal Control Deficiencies and Their Remediation on Accrual Quality. *The Accounting Review*, *83*(1), 217–250.

Asthana, S., & Boone, J. P. (2012). Abnormal Audit Fee and Audit Quality. *Auditing: A Journal of Practice & Theory*, *31*(3), 1–22.

Bae, G. S., Choi, S. U., Lamoreaux, P. T., & Lee, J. E. (2021). Auditors' Fee Premiums and Low-Quality Internal Controls. *Contemporary Accounting Research*, *38*(1), 586–620.

Ball, R., & Shivakumar, L. (2006). The Role of Accruals in Asymmetrically Timely Gain and Loss Recognition. *Journal of Accounting Research*, *44*(2), 207–242.

Blankley, A. I., Hurtt, D. N., & MacGregor, J. E. (2012). Abnormal Audit Fees and Restatements. *Auditing: A Journal of Practice & Theory*, *31*(1), 79–96.

Blankley, A. I., Hurtt, D. N., & MacGregor, J. E. (2013). Evaluating the Effect of Abnormal Audit Fees on Future Restatements. *Current Issues in Auditing*, *7*(1), 15–21.

Blankley, A. I., Hurtt, D. N., & MacGregor, J. E. (2014). The Relationship between Audit Report Lags and Future Restatements. *Auditing: A Journal of Practice & Theory*, *33*(2), 27–57.

Calderon, T. G., Wang, L., & Klenotic, T. (2012). Past control risk and current audit fees. *Managerial Auditing Journal*, *27*(7), 693–708.

Cao, Y., Myers, L. A., & Omer, T. C. (2012). Does Company Reputation Matter for Financial Reporting Quality? Evidence from Restatements. *Contemporary Accounting Research*, *29*(3), 956–990.

Casterella, J. R., Francis, J. R., Lewis, B. L., & Walker, P. L. (2004). Auditor Industry Specialization, Client Bargaining Power, and Auditing Pricing. *Auditing: A Journal of Practice & Theory*, *23*(1), 123–140.

Chalmers, K., Hay, D., & Khlif, H. (2019). Internal control in accounting research: A review. *Journal of Accounting Literature*, *42*, 80–103.

Chen, H., Zhang, M., & Ling, H. (2011). Restatement and Audit Risk. *Financial Decisions*, *23*(1), 1–14.

Chen, R. S. (2019). Internal Control Audit Fee and Internal Control Audit Quality - Evidence from Integrated Audits. *Open Journal of Business and Management*, *7*, 292–311.

Chen, Y., Smith, A. L., Cao, J., & Xia, W. (2014). Information Technology Capability, Internal Control Effectiveness, and Audit Fees and Delays. *Journal of Information Systems*, *28*(2), 149–180.

Choi, J.-H., Kim, J.-B., & Zang, Y. (2010). Do Abnormally High Audit Fees Impair Audit Quality? *Auditing: A Journal of Practice & Theory*, *29*(2), 115–140.

Committee of Sponsoring Organizations of the Treadway Comission. (2012). Internal Control - Integrated Framework, 1–178.

DeFond, M., & Zhang, J. (2014). A review of archival auditing research. *Journal of Accounting and Economics*, *58*, 275–326.

Doogar, R., Sivadasan, P., & Solomon, I. (2015). Audit fee residuals: Costs or rents? *Review of Accounting Studies*, *20*, 1247–1286.

Dopuch, N., Mashruwala, R., Seethamraju, C., & Zach, T. (2012). The Impact of a Heterogeneous Accrual-Generating Process on Empirical Accrual Model. *Journal of Accounting, Auditing & Finance*, *27*(3), 386–411.

Doyle, J., Ge, W., & McVay, S. (2007). Determinants of weaknesses in internal control over financial reporting. *Journal of Accounting and Economics*, *44*(1-2), 193–223.

Elder, R., Zhang, Y., Zhou, J., & Zhou, N. (2009). Internal Control Weaknesses and Client Risk Management. *Journal of Accounting, Auditing & Finance*, *24*(4), 543–579.

Eshleman, J. D., & Guo, P. (2014). Abnormal Audit Fees and Audit Quality: The Importance of Considering Managerial Incentives in Tests of

[165] See therefore Pittman and Zhao (2021, p. 129).

[166] See therefore Albring et al. (2018, p. 485); H. Chen et al. (2011, p. 2); Hribar and Jenkins (2004, p. 337); Palmrose et al. (2004, p. 59).

Earnings Management. *Auditing: A Journal of Practice & Theory*, *33*(1), 117–138.

Fang, L., & Yasuda, A. (2009). The effectiveness of reputation as a disciplinary device in sell-side research. *Review of Financial Studies*, *22*(9), 3735–3777.

Feldmann, D. A., Read, W. J., & Abdolmohammadi, M. J. (2009). Financial Restatements, Audit Fees, and the Moderating Effect of CFO Turnover. *Auditing: A Journal of Practice & Theory*, *28*(1), 205–223.

Fields, L. P., Fraser, D. R., & Wilkins, M. S. (2004). An investigation of the pricing of audit services for financial institutions. *Journal of Accounting and Public Policy*, *23*, 53–77.

Gaber, M., Garas, S., & Lusk, E. J. (2019). Evidence on the impact of internal control over financial reporting on audit fees. *International Journal of Accounting and Financial Reporting*, *9*(3), 1–23.

Garcia Osma, B. (2008). Board Independence and Real Earnings Management: The Case of R&D Expenditure. *Corporate Governance: An International Review*, *16*(2), 116–131.

Guragai, B., & Hutchinson, P. D. (2019). Material weakness disclosures and restatements: value of external auditor attestation. *Accounting Research Journal*, *32*(3), 362–380.

Hammersley, J. S., Myers, L. A., & Shakespeare, C. (2008). Market reactions to the disclosure of internal control weaknesses under Section 302 of the Sarbanes Oxley Act of 2002. *Review of Accounting Studies*, *23*(1), 141–191.

Hammersley, J. S., Myers, L. A., & Zhou, J. (2012). The Failure to Remediate Previously Disclosed Material Weaknesses in Internal Controls. *Auditing: A Journal of Practice & Theory*, *31*(2), 73–111.

Hay, D. (2013). Further Evidence from Meta-Analysis of Audit Fee Research. *International Journal of Auditing*, *17*, 162–176.

Hoag, M. L., & Hollingsworth, C. W. (2011). An Intertemporal Analysis of Audit Fees and Section 404 Material Weaknesses. *Auditing: A Journal of Practice & Theory*, *30*(2), 173–200.

Hogan, C. E., & Wilkins, M. S. (2008). Evidence on the Audit Risk Model: Do Auditors Increase Audit Fees in the Presence of Internal Control Deficiencies? *Contemporary Accounting Research*, *25*(1), 219–242.

Hoitash, R., Hoitash, U., & Bedard, J. C. (2008). Internal Control Quality and Audit Pricing under the Sarbanes-Oxley Act. *Auditing: A Journal of Practice & Theory*, *27*(1), 105–126.

Hoitash, R., Markelevich, A., & Barragato, C. A. (2007). Auditor fees and audit quality. *Managerial Auditing Journal*, *22*(8), 761–786.

Hribar, P., & Jenkins, N. T. (2004). The Effect of Accounting Restatement on Earnings Revisions and the Estimated Cost of Capital. *Review of Accounting Studies*, *9*(2-3), 337–356.

Hribar, P., Kravet, T., & Wilson, R. (2014). A new measure of accounting quality. *Review of Accounting Studies*, *19*, 506–538.

Jackson, A. R. (2005). Trade Generation, Reputation, and Sell-Side Analysts. *The Journal of Finance*, *60*(2), 673–717.

Jiang, W., & Son, M. (2015). Do Audit Fees Reflect Premiums for Control Risk? *Journal of Accounting, Auditing & Finance*, *30*(3), 318–340.

Jonas, G., Rosenberg, A., Gale, M., & Hedges, L. (2007). The Third Year of Section 404 Reporting on Internal Control.

Jones, J. J. (1991). Earnings Management During Import Relief Investigations. *Journal of Accounting Research*, *29*(2), 193–228.

Kester, L., Georgakopoulos, G., Kalantonis, P., & Boufounou, P. (2013). The Impact of Board of Directors' Turnover on the Association between Financial Restatements and Audit Fees. *Journal of Applied Finance & Banking*, *3*(2), 153–175.

Krauß, P., Pronobis, P., & Zülch, H. (2015). Abnormal audit fees and audit quality: initial evidence from the German audit market. *J Bus Econ*, *85*, 45–84.

Krishnan, J., Krishnan, J., & Song, H. (2011). The Effect of Auditing Standard No. 5 on Audit Fees. *Auditing: A Journal of Practice & Theory*, *30*(4), 1–27.

Lawrence, A., Minutti-Meza, M., & Vyas, D. (2018). Is Operational Control Risk Informative of Financial Reporting Deficiencies? *Auditing: A Journal of Practice & Theory*, *37*(1), 139–165.

Lobo, G. J., & Zhao, Y. (2013). Relation between Audit Effort and Financial Report Misstatements: Evidence from Quarterly and Annual Restatements. *The Accounting Review*, *88*(4), 1385–1412.

Mitra, S. (2009). Pervasiveness, severity, and remediation of internal control material weaknesses under SOX Section 404 and audit fees. *Review of Accounting and Finance*, *8*(4), 369–387.

Mitra, S., Jaggi, B., & Al-Hayale, T. (2017). The effect of managerial stock ownership on the relationship between material internal control weaknesses and audit fees. *Review of Accounting and Finance*, *16*(2), 239–259.

Munsif, V., Raghunandan, K., Rama, D. V., & Singhvi, M. (2011). Audit Fees after Remediation of Internal Control Weaknesses. *Accounting Horizons*, *25*(1), 87–105.

Palmrose, Z.-V., Richardson, V. J., & Scholz, S. (2004). Determinants of market reactions to restatement announcements. *Journal of Accounting and Economics*, *37*, 59–89.

PCAOB. (2007). Public Company Accounting Oversight Board, PCAOB Release No. 2007-005A.

PCAOB. (2010). Public Company Accounting Oversight Board, PCAOB Release No. 2010-004.

PCAOB. (2013). Public Company Accounting Oversight Board, Considerations for Audits of Internal Control over Financial Reporting. Staff Audit Practice Alert No. 11.

Pittman, J. A., & Zhao, Y. (2021). Auditor Responses to and Prevention of Non-Income-Increasing Misreporting: Evidence from Audit Fees and Restatements. *Auditing: A Journal of Practice & Theory*, *40*(3), 127–153.

Raghunandan, K., & Rama, D. V. (2006). SOX Section 404 Material Weakness Disclosures and Audit Fees. *Auditing: A Journal of Practice & Theory*, *25*(1), 99–114.

Rajgopal, S., Srinivasan, S., & Zheng, X. (2021). Measuring audit quality. *Review of Accounting Studies*, *26*, 559–619.

Romanus, R. N., Maher, J. J., & Fleming, D. M. (2008). Auditor Industry Specialization, Auditor Changes, and Accounting Restatements. *Accounting Horizons*, *22*(4), 289–413.

Seidel, T. A. (2017). Auditors' Response to Assessments of High Control Risk: Further Insights. *Contemporary Accounting Research*, *34*(3), 1340–1377.

Simunic, D. A. (1980). The Pricing of Audit Services: Theory and Evidence. *Journal of Accounting Research*, *18*(1), 161–190.

Stein, M. T., Simunic, D. A., & O'Keefe, T. B. (1994). Industry Differences in the Production of Audit Services. *Auditing: A Journal of Practice & Theory*, *13*(Supplement), 128–142.

United States Government Accountability Office. (2006). Financial Restatements: Update of Public Company Trend, Market Impacts, and Regulatory Enforcement Activities [Report to the Ranking Minority Member, Committee on Banking, Housing, and Urban Affairs, U.S. Senate, 1-206].

Viciwati. (2020). Bankruptcy Prediction Analysis using the Zmijewski Model (X-Score) and the Altman Model (Z-Score). *Dinasti International Journal of Economics, Finance & Accounting*, *1*(5), 794–806.

Zmijewski, M. E. (1984). Methodological Issues Related to the Estimation of Financial Distress Prediction Models. *Journal of Accounting Research*, *22*, 59–82.

Junior Management Science 10(3) (2025) 677-714



Junior Management Science

www.jums.academy
ISSN: 2942-1861



Mandatory ESG Disclosure and Firm Value – A Quantitative Analysis of the Effect of Directive 2014/95/EU on Firm Value

Jan Oliver Horstmann

WHU - Otto Beisheim School of Management

Abstract

This work offers novel insights into how the introduction of the Non-Financial Reporting Directive in the European Union in 2014 affected firm value. Based on the theoretical discourse, it seems ex-ante unclear how this ESG disclosure directive is perceived by capital markets and affects firm value. Hence, this work aims to shed more light on the topic and add to the scant evidence literature offers. Specifically, the implications of the first-time mandate of ESG information disclosure are investigated using an instrumental variable and a difference-in-difference approach on a propensity score-matched sample of 708 firms based in the European Union and the U.S. Difference-in-difference results imply that firm's ESG performance, measured by Refinitiv's ESG scores, significantly increases after the adoption of the directive. Subsequent instrumental variables analysis suggests that the increased ESG performance resulting from the directive is associated with relatively weak, negative effects for Tobin's Q as the measure of firm value. In addition to confirming anticipatory effects for Tobin's Q as early as 2014, significant evidence reveals that firms (sectors) with higher ESG performance had a more negative market reaction than firms (sectors) with lower ex-ante ESG performance.

Keywords: disclosure regulation; firm value; mandatory ESG reporting; market reaction; non-financial reporting directive (NFRD)

1. Introduction

"The Social Responsibility of Business Is to Increase Its Profits" (Friedman, 1970)

Following this narrow neoclassical economic view that Milton Friedman voiced in an essay in the *New York Times*, managers should only engage in business activities that maximise shareholder value. Friedman labelled it "immoral" of managers to engage in Corporate Social responsibility (CSR) activities and that this should be considered "stealing" from shareholders (Freeman & Dmytriyev, 2017; Friedman, 1970; Kitzmueller & Shimshack, 2012). Pointedly depicting the last 50 years of political and economic discussion about this controversial paradigm, the New York Times published 22 responses from CEOs, economists, and Nobel laureates to the "Friedman´s doctrine" on the 50^{th} anniversary of Friedman´s essay. Most of the former opposed Friedman's views and pleaded for a more stakeholder-oriented behaviour in business (Sorkin, 2020). Part of this move from Shareholder- to Stakeholder-value-centrism is the increasing focus on externalities that are neglected within a narrow interpretation of Friedman's paradigm (Liang & Renneboog, 2020; Magill et al., 2015).

With the plethora of interwoven problems such as climate change, hunger, and migration that the world faces, stakeholder-centrism has been more significant than ever before, as from a societal perspective, internalising externalities and distributing rights and assets across generations can support approaching these problems (Christensen et al., 2018; Howarth & Norgaard, 1992; Magill et al., 2015). This is why, in the last decades, integrating ESG[1] criteria in investment decisions has become a new normal. As the PRI reported, by March 2022, 4,902 investors and service

[1] Environmental, Social, and Governance; see Chapter 2.1 for further definition on the terminology.

DOI: https://doi.org/10.5282/jums/v10i3pp677-714

© The Author(s) 2025. Published by Junior Management Science.
This is an Open Access article distributed under the terms of the CC-BY-4.0 (Attribution 4.0 International). Open Access funding provided by ZBW.

providers representing US$121.3 trillion of assets under management committed to incorporating ESG factors into their investment decisions (Principles for Responsible Investment, 2022). Along with the over the years steadily increasing demand for investing sustainably, the resulting demand for information on companies' ESG performance is growing as well (Amel-Zadeh & Serafeim, 2017; Berg et al., 2022; Eccles et al., 2011). Therefore, a resulting question will not evolve around whether or not to engage in CSR activities but if disclosure of information on ESG prompts value creation (Liang & Renneboog, 2020).

Responding to the increased focus on ESG criteria in investment decisions and the resulting demand for information on ESG performance, an augmented number of companies voluntarily disclose this information. For example, from 2010 to 2022, the number of S&P 500 companies voluntarily publishing ESG reports grew from 20% to 96% (G&A, 2021, 2023). In the U.S., afar from financial, risk, and litigation disclosure requirements concerning public companies and environmental issues, as well as human capital management-related aspects from 2020 onwards, no mandatory requirements on disclosing ESG information have been laid out (Cifrino, 2023). In comparison, many jurisdictions, such as the European Union (EU), have already introduced mandatory reporting mandates (Ioannou & Serafeim, 2017).

In 2014, the EU adopted the first directive to foster sustainability-related information disclosure: The Non-Financial Reporting Directive (2014/95/EU; NFRD)[2], which came into force in 2017. The NFRD, transitioning ESG reporting in the EU from voluntary to mandatory, requires large companies with more than 500 employees and more than EUR€ 20,000,000 of total assets or revenues of more than EUR€ 40,000,000 to publish information related to environmental matters, social matters and treatment of employees, respect for human rights, anti-corruption and bribery, and diversity on company boards (European Commission, 2014a).

Despite a growing body of research agreeing that voluntary and mandatory ESG disclosure affects firm value, current literature neither theoretically nor empirically provides a clear answer to whether this effect is positive or negative.[3] Research in favour of a firm value-increasing effect argues that, e.g., elevated ESG performance would lead to positive valuation effects (Bajic & Yurtoglu, 2018; Fatemi et al., 2018; Flammer, 2015; Ioannou & Serafeim, 2017) while opposing literature brings forward, e.g., that the mandated disclosure could incur higher costs for companies that would need to distinguish themselves from competing firms (Grewal et al., 2019; Ioannou & Serafeim, 2017). Not only is literature regarding firm value implications of ESG disclosure mandates inconsistent in its findings, it also is relatively scant and more recent (Christensen et al., 2021; Hummel & Jobst, 2022; Mittelbach-Hörmanseder et al., 2021). To support closing this research gap, the present study aims to answer the research question of how the introduction of the NFRD affected firm value.

The supranational and regulation-driven nature of this introduction across the EU presents an ideal setting to exploit the effect of mandatory ESG disclosure on firm value in a shock-based instrumental variables model (referred to as "IV" in the following) (Atanasov & Black, 2016; Christensen et al., 2017; La Torre et al., 2018; Leuz & Wysocki, 2015; Mittelbach-Hörmanseder et al., 2021). Specifically, the primary model comprises a two-staged least squares instrumental variables model (referred to as "2SLS" in the following) with a difference-in-difference Ordinary Least Squares (OLS) regression (referred to as "DID" in the following) in the first stage. In this first stage, the direct effect of the NFRD on the dependent variable ESG performance (measured as ESG score from Refinitiv) is predicted. In the second stage, the predicted values from the first stage are employed to predict the outcome of Tobin's Q. In addition to clustered standard errors and firm, year and sector-year fixed effects, various control variables such as size and leverage are included in the models. This model is run for the three subcomponent pillars of ESG as well. Furthermore, yearly effects are estimated in a two-staged DID model, and cross-sectional partitions are introduced based on sectoral affiliation and pre-directive ESG performance.

Five predictions are made to answer the research question of how the NFRD affected firm value. First, an on-average negative firm value reaction to the NFRD is hypothesised. Second, the reaction for prediction one will likely materialise differently depending on the topic concerned (Environmental, Social, or Governance). Third, this firm value reaction is likely more significant in the late post-period (2018-2019) than in the early post-period (2014-2017) of the directive. Fourth, in a cross-sectional partition, the effect of prediction one is likely to be more (less) significant for firms with lower (higher) pre-directive ESG performance. Fifth, entangled with prediction fours, sectoral affiliation is likely to be of relevance to firm value implications.

The statistically significant findings provide evidence that, on average, a firm's ESG performance increases after the adoption of the directive. Succeeding instrumental variables analysis utilising the predicted ESG performance from the first stage indicates that the positive and significant ESG performance is associated with negative effects for Tobin's Q as the measure of firm value. As yearly effects estimation indicates, this negative firm value outcome is most pronounced for the year of the directive's adoption, 2014, while it is 2015 for the ESG performance. Thus, demonstrating the existence of anticipatory effects after the adoption of the regulation. Additionally, the findings of this work corroborate the value-relevance of sector affiliation and pre-directive ESG performance of firms subject to the mandated disclosure. Statistically significant evidence brought forward in this work reveals that based on the sample, firms with

[2] Referred to as „NFRD", „the directive", or „the mandate" in the following.

[3] Despite the mixed literature, most of the empirical evidence implies a positive association between voluntary disclosure of ESG information and firm value. This indication needs to be treated, as mentioned, with carefulness due to the dual-selection problem (Christensen et al., 2021; Hummel & Jobst, 2022; Mittelbach-Hörmanseder et al., 2021).

higher ESG performance or those who belong to sectors with an on average high ESG rating had a more negative market reaction than firms with lower ex-ante ESG performance or those who belonged to lower ESG-rated sectors.

This work makes numerous research contributions. First, to the best of current knowledge, this study is the first to analyse the impact of the NFRD on firm value (proxied by Tobin's Q) in an instrumental variables setting, supported by results from difference-in-difference analyses. In addition, it is the first work that examines firm value consequences of the NFRD in a yearly effects model. As mentioned, this work expands limited research on the value implications of the NFRD and lends weight to the prevalent understanding that the NFRD had an overall negative effect on firm value (Fiechter et al., 2022; Grewal et al., 2019; Mittelbach-Hörmanseder et al., 2021). Besides, scholarship on the anticipatory effects of mandatory disclosure regulation (Fiechter et al., 2022; Grewal et al., 2019) as well as the relevance of sectoral affiliation (Cahan et al., 2016; Christensen et al., 2017; Eccles et al., 2012; Krueger, 2015a) and prior ESG performance (Fiechter et al., 2022; Grewal et al., 2019; Ioannou & Serafeim, 2017; Jackson et al., 2020; Mittelbach-Hörmanseder et al., 2021) is enriched by the insights generated in this work. As ESG disclosure regulation is a central topic of national and international interest, the findings of this work hold significance for regulators, policymakers, and firms alike.

The scope of this work is limited in various ways. This work mainly focuses on the overall effects of the NFRD on firm value and does not examine through which channels these effects materialise or why firms engage in the underlying ESG activities. Furthermore, it is pertinent to mention that this work analyses associations and not causal relationships and cannot be not entirely immune to the influence of flawed data (e.g., through greenwashing) provided by Refinitiv.

The remainder of this work is structured as follows. Chapter 2. provides background knowledge on mandatory ESG information disclosure and the theoretical concepts behind it and elucidates explicitly on the introduction of the NFRD in Europe. Chapter 3. offers a comprehensive overview of the literature related to mandatory ESG disclosure and its relationship with firm value. Based on the theoretical background and the literature review, the hypotheses 1. to 3. are developed in this chapter. Subsequently, Chapter 4. gives an overview of the sample construction and the main variables used in the empirical analysis. Chapter 5. follows with the research design to test the hypotheses to answer the research question. Results of the empirical analysis and robustness tests are reported in Chapter 6. This work concludes with Chapter 7. which summarises the findings and elaborates on the contributions, implications, and limitations and points out further research potential.

2. Background

2.1. Relevant Definitions

Since this work deals with the impact of the NFRD on firm value, clarifying the terminology related to the NFRD is indispensable to developing, analysing, and evaluating a thorough research agenda on the implications of such a policy intervention.

Terms such as "Sustainability", "Corporate Sustainability", "Corporate Citizenship", "Corporate Social Responsibility" (CSR), "non-financial", or "Environmental – Social – Governance" (ESG) pertaining to information, encompass terms that are related but different. The scientific literature still has not conclusively determined clear boundaries between the previous terms because of the ambiguity that is associated with the concepts (Christensen et al., 2021; Dhaliwal et al., 2011; Ioannou & Serafeim, 2017; Park et al., 2023; Sheehy & Farneti, 2021; Tarquinio & Posadas, 2020).

An example of this is the definition of a concept of non-financial information. As a shared agreement on this definition is missing, a definition of non-financial is still open to interpretation. Following a systematic literature review by Tarquinio and Posadas (2020), scholars mainly define this broad term with a residual definition, thus not explaining what non-financial information is but rather what it is not. Tarquinio and Posadas's (2020, p. 743) research highlights that non-financial information can also pertain to "intellectual capital information, strategy, business performance and risk." which may concern other areas than CSR, ESG, and sustainability. The authors conclude that the term non-financial, concerning disclosed information, forms a "genus", an umbrella term that can overlap and be synonymous with other information concepts such as ESG, CSR and Sustainability (Tarquinio & Posadas, 2020).

In a similar fashion, the terms CSR and ESG are interchangeably used as well. Both are the most predominant terms in the relevant literature, published by respected scholars in established journals consulted to create this research. (Christensen et al., 2021; Dhaliwal et al., 2011; Grewal et al., 2019; Ioannou & Serafeim, 2017; Liang & Renneboog, 2020; Mittelbach-Hörmanseder et al., 2021).

There are various definitions of CSR. Christensen et al. (2021, p. 1181) define CSR as "... corporate activities and policies that assess, manage, and govern a firm's responsibilities for and its impacts on society and the environment". CSR includes a wide range of activities that can be subsumed under the term "ESG". The acronym for "Environmental", "Social", and "Governance" is most commonly used in the capital markets environment, as the more clear division into three pillars makes it easier to distinguish, assign, and evaluate performance (Berg et al., 2022; Christensen et al., 2018; Harper Ho, 2016). For reasons of clarity and brevity, the terms ESG and CSR will be used interchangeably in the following work. Nevertheless, the term ESG will be used more frequently per its prior definition as the following work mainly examines non-financial disclosure in a capital markets environment. Moreover, in recent years, the term ESG

has gained increased traction in news and literature and will likely keep growing in importance (Christensen et al., 2018; Eccles et al., 2011; Pollman, 2021). A Google trends comparison (web search and news search) of the terms CSR and ESG can support this statement (*Appendix F;* Google Trends, 2023).

2.2. Non-financial Reporting in the European Union

With the supranational "Non-financial Reporting Directive" (NFRD; Council Directive 2014/95/EU), ESG disclosure was mandated by the EU for the first time. Driven by the goal to improve ESG disclosure of specific companies, on April 16, 2013, the European Commission proposed an amendment to the existing Accounting Directive (Directive 2013/34/EU). In the non-financial reporting realm, this directive set out requirements concerning policies, risks and outcomes linked to environmental and employee-related matters in annual financial statements (Allman & Won, 2022; Costa & Agostini, 2016). The NFRD was agreed on by the European Parliament and Council on February 26, 2014, and was adopted by the Parliament on April 15, 2014. Following Article 288 of the Treaty on the Functioning of the European Union, the member states were left to some degree of independence in the transposition of the directive into national law. The directive required the member states to adopt the directive into national law by the end of 2016. The first reporting year for companies to apply the directive was set to be the financial year starting on January 1, 2017, or during the calendar year 2017. Followingly, first-time reporting companies were mandated to publish the information in 2017 (Agliati, 2021; Allman & Won, 2022; European Commission, 2014a). All relevant events around the NFRD are tabulated in *Appendix C.*

The Directive 2014/95/EU (2014a) requires "large undertakings which are public-interest entities exceeding on their balance sheet dates the criterion of the average number of 500 employees during the financial year shall include in the management report a non-financial statement containing information to the extent necessary for an understanding of the undertaking's development, performance, position and impact of its activity, relating to, as a minimum, environmental, social and employee matters, respect for human rights, anti-corruption and bribery matters" (Article 19a (1); for further information on the scope of the directive, see Directive 2014/95/EU). With the included "double materiality" principle, companies are forced to disclose how ESG matters may affect the company's future and how the company affects society. Moreover, companies can decide not to implement policies pertaining to the matters mentioned in Article 19a (1). If they do so, companies "shall provide a clear and reasoned explanation for not doing so" (*ibidem*). This is referred to as the "comply-or-explain-approach". Not only do companies not need to follow a specific reporting guideline or framework for their disclosure, but they can – under specific requirements – publish a standalone ESG report (Article 1 (4)). The disclosures are to be checked in a binary manner by external auditors to determine whether they have been included. Even if there is no legally binding obligation to verify the contents of the disclosure, transposing countries can decide if they want the reported information to be verified by an external auditor (Article 1 (5); (6)).

A definition for "large undertakings that are public-interest entities" can be found in Article 2 of Directive 2013/34/EU (2013). Public interest entities include listed companies, credit institutions, insurance companies, and other companies "whose transferable securities are admitted to trading on a regulated market of any Member State" (Article 2 (1)). A company is a large undertaking if it fulfils two of the three requirements laid out in Article 3 (4): Balance sheet total of EUR€ 20,000,000, net turnover[4] of EUR€ 40,000,000, average number of employees during the financial year: 250 (European Commission, 2013).

Following the minimum thresholds, and with subsidiaries of other public interest entities exempted from the reporting mandate, the number of companies required to report under the NFRD sums to 1,956 (0.02% of all EU27 LLCs with 27% of total turnover of all EU LLCs and 86% market cap). Since member states are given the freedom to apply directive 2014/95/EU and 2013/354/EU to a broader set of firms in their national transposition, 13 states have introduced a wider definition of the minimum threshold, ramping up the number of mandated companies to around 11,500 (De Groen et al., 2020). Before the introduction of the NFRD, only around 2,500 larger undertakings disclosed ESG information regularly (European Commission, 2014b).

Since 2014, the EU has adopted five significant non-financial reporting legislations. The NFRD, the Capital Requirements Regulation II (CRR; Pillar 3 disclosure; publication on June 7, 2019), the Sustainable Finance Disclosure Regulation (SFDR; publication on December 9, 2019), the Taxonomy Regulation (publication on June 22, 2020), and the Corporate Sustainability Reporting Directive (CSRD; publication on December 16, 2022) (European Commission, 2019a, 2019b, 2020, 2022).

2.3. Principles of Mandatory ESG Information Disclosure in the Context of Firm Value

Simultaneous with the rise in interest in ESG information, regulation grew and, hence, the scientific literature on mandating non-financial information. In a study, Fiandrino et al. (2022) found that from 2016 onwards, nearly 60% of scientific literature on non-financial information was published between 2019 and 2021. Moreover, scientific literature output in this field also grew in absolute numbers during that time. Despite this increase in literature on non-financial information regulation in the past years, the prevalent arm of research in corporate information disclosure and

[4] Net turnover is defined by Directive 2013/34/EU (2013) as "the amounts derived from the sale of products and the provision of services after deducting sales rebates and value added tax and other taxes directly linked to turnover" (Article 2 (5)).

its regulation still concentrates on *financial* reporting and disclosure regulation. Within academic research on financial and non-financial information disclosure, scholars differentiate between mandatory and voluntary disclosure (Krueger, 2015a; Leuz & Wysocki, 2015).

The difference between mandatory and voluntary ESG disclosure research exists in the partly divergent theoretical basis for both. Opposed to the scientific discourse on mandatory reporting, the discourse on voluntary reporting of ESG information entails looking at, e.g., adoption decisions of individual subjects or reporting-standard adherence. These adoption decisions need not be undertaken by firms in a mandatory setting, but the firm's underlying considerations will play a role in hypothesizing the effect of the NFRD on firm value. (Christensen et al., 2018, 2021; Mittelbach-Hörmanseder et al., 2021). Since this work deals with the link between the NFRD and firm value, only theoretical economic and socio-political concepts related to this regulation, needed to draw profound assumptions and later on assess the potential impact of the NFRD on firm value, will be discussed for reasons of clarity and briefness.

ESG disclosure can have many economic consequences; not all can be mentioned in this concise theoretical overview. In the presence of the NFRD, however, from a conceptual point of view, the theoretical discussion about mandatory ESG disclosure begins with the two opposing "shareholder expense view" and the "stakeholder maximization view" (Grewal et al., 2019; Leuz & Wysocki, 2008; Manchiraju & Rajgopal, 2017; Mittelbach-Hörmanseder et al., 2021). The discussion will be expanded by thoughts from a framework proposed by Leuz and Wysocki (2008), who approach disclosure and reporting regulation from a firm-specific (microeconomic) and market-wide (macro-economic) angle.

Taking up Friedman's (1970) shareholder expense view from the introduction of this work, it follows that CSR activities have a negative influence on firm value as a firm engages in CSR activities at the expense of shareholders (Mittelbach-Hörmanseder et al., 2021). Costs associated with this include direct costs and indirect costs. Direct costs resulting from a mandated ESG disclosure regime stem from complying with the NFRD (e.g. preparing the report or adapting internal processes). As Grewal et al. (2019) point out, these costs are relatively lower than others. An ex-ante assessment of the cost to comply with the NFRD resulted in an average administrative cost of EUR€ 82,000 per firm per year (De Groen et al., 2020). The political and proprietary costs are more significant for assessing the impact of the NFRD on firm value. Both form the indirect cost part. Proprietary costs refer to a potential decrease in the competitiveness of a firm subject to the mandate and thus influence its current and future attractiveness for investors. A firm will moreover carry costs if it decides to maintain its weak ESG performance and, e.g., pay for penalties or if it decides to improve its ESG performance (Chang et al., 2022; Christensen et al., 2021; Grewal et al., 2019). Political costs relate to current and priced-in future costs stemming from the politically induced pursuit of an investment, which is perceived as an investment carrying a negative net present value to shareholders. Actors asserting pressure can be the government, the regulator, or other interest groups, whether governmental or non-governmental. (Chang et al., 2022; Grewal et al., 2019; Healy & Palepu, 2001; Jensen & Meckling, 1978; Watts & Zimmerman, 1978).

In contrast to the cost that a firm is confronted with when disclosing ESG information, the stakeholder maximization view posits that benefits can come along with it as well, and firm value can be positively affected (Deng et al., 2013; Mittelbach-Hörmanseder et al., 2021). Analogue to the costs, the benefits can again be split into capital market effects and real effects, which concern the behaviour of the disclosing firm. Both effects are interconnected and relevant for the implications of the NFRD for firm value, as capital market effects are a tangible visualization of the expected real effects (e.g., benefits) for the disclosing firm (Leuz & Wysocki, 2015). From an investor point of view, the most significant benefit of mandatory disclosure of ESG information is the reduction of information asymmetries between a firm and potential investors as well as within a group of investors (Christensen et al., 2021; Verrecchia, 2001). First, mitigating the adverse selection problem leads to a lowered return threshold at which investors may be willing to invest in a firm (Amihud & Mendelson, 1986; Constantinides, 1986). Second, the information risk for an investor is lowered. The investor can have a more informed approach to valuing a firm, predicting a firm's future potential, and assessing a firm's risks, hence reducing the firm's cost of capital (Easley & O'Hara, 2004; Grewal et al., 2019; Krueger et al., 2021; Lambert et al., 2012). Third, monitoring effectiveness, e.g., observing environmental performance, is increased and, thus, corporate investment efficiency alike (Krueger et al., 2021; Lambert et al., 2007). Fourth, mandated ESG disclosure could advance the operational efficiency of a firm by forcing them to improve product quality, reduce energy consumption and waste and enhance the firm's reputation, brand value, or employee retention, among other reasons (Chang et al., 2022; Christensen et al., 2021; Grewal et al., 2019; Ioannou & Serafeim, 2017).

If only approached from a (net) benefit-to-the-firm view, following the full-disclosure-equilibrium, disclosure can be kept in a voluntary setting, and no need for regulatory intervention is given, as Leuz and Wysocki (2008) point out in detail. Conversely, considering the existence of firm-specific and market-wide costs and benefits, the overall equation may favour mandatory ESG disclosure even if firm-specific net benefits are negative, as higher market-wide net benefits can compensate for them. Moreover, even if market-wide, beneficial effects are significant, in most cases, these effects cannot be internalized by companies which would thus not disclose ESG information voluntarily. Hence, mandatory disclosure can help provide an optimal level of ESG disclosure (Christensen et al., 2021; Leuz & Wysocki, 2008).

As mentioned, market-wide net benefits are not internalized by firms and, thus, play a minor role in assessing the effects of the NFRD on firm value. Nevertheless, they

were still mentioned as they are crucial in understanding the motivation for mandating ESG disclosure in the European Union. Assessing the aggregated effect of the NFRD on firm value implies the need to hypothesize whether the aggregated economic effect is positive or negative. The effect is the sum of the net benefit of every firm expected by the market. This individual net benefit can vary depending on various factors outlined in the previous chapter as well as existing pre-directive reporting on ESG, sector affiliation, and further (Grewal et al., 2019; Mittelbach-Hörmanseder et al., 2021). It has to be mentioned that a potential negative outcome in firm value does not necessarily imply that aggregated economic value has been destroyed. Moreso, embedded in an extended economic analysis, this work can be part of an overall analysis of the benefits of the NFRD for society.

2.4. Background on Mandatory ESG Information Disclosure

Decades of environmental and social activism preceded the world's first mandatory ESG reporting mandates. It was only in the 60's and 70's of the past century that companies were increasingly held accountable for their negative environmental and social impact. For example, in the 1960s, Vietnam War protestors demanded university endowment funds to step down from investments in the defence sector (Oreskes, 1985). The first attempts at voluntary social reporting were made in Europe by France and the Netherlands, swiftly followed by Austria, Switzerland, and Germany. In the 1980s – from a more religiously motivated impetus – ethical investment funds from the United Kingdom and the United States began basing investment decisions on a "negative screening" approach. This excluded companies from their investment universe that engaged in business labelled as "sinful", such as gambling, alcohol, and tobacco. In the 1990s, the trend for more ESG focus in investment decisions grew more substantially due to increased societal pressure following various environmental disasters and media coverage over the consequences of globalization (Ioannou & Serafeim, 2017; Kolk, 2003). As a result, the United Nations Environment Program (UNEP) and the Coalition for Environmentally Responsible Economies (CERES) jointly created the Global Reporting Initiative (GRI) in 1997, which provides the world's most widely used sustainability reporting standard covering each E, S, and G pillar (GRI, 2022).

Around the time of the Global Financial Crisis, investors and the public increasingly demanded reporting on ESG matters. Following these demands, a growing number of countries began mandating ESG information disclosure in the early 2000's. Countries such as France (2001), Australia (2003), Canada (2004), Malaysia (2007), Denmark (2008), China (2008), or South Africa (2010) all issued various mandatory ESG disclosure regulations around that time (Krueger et al., 2021). A landmark event in the century following was the Paris Agreement on Climate Change in 2015, as well as the adoption of the Sustainable Development Goals (SDGs) by the United Nations in the same year. Both events pointedly put climate change and sustainability into the spotlight of the world stage (Hummel & Jobst, 2022). As of 2020, the Carrots and Sticks[5] database listed 1,102 active, mandatory ESG policies in 84 countries worldwide (Van der Lugt et al., 2020).

In the context of ESG disclosure mandates, the United States (U.S.) remain a particular case. Until now, afar from financial, risk, and litigation disclosure requirements concerning public companies with environmental issues and human capital management-related aspects from 2020 onwards, no mandatory requirements on disclosing ESG information have been introduced (Cifrino, 2023). This is one reason why the U.S. represents a clean control group to assess the impact of the NFRD. According to the Governance and Accountability Institute (G&A), the number of S&P 500 companies voluntarily publishing ESG reports grew from 20% to 96% from 2010 to 2022 (G&A, 2021, 2023). In March 2022, the U.S. Securities and Exchange Commission proposed a novel mandate requiring registered foreign or domestic firms to include climate-related information, from GHG emissions to climate risks, in registration statements and reports such as the 10-K annual report (S&P Global, 2023). In April 2023, the SEC updated the status of the mandate, with most agenda points showing finalization dates between October 2023 and April 2024 (U.S. General Services Administration, 2023).

3. Literature Review and Hypotheses Development

3.1. Prior Research on the Relationship of Mandatory ESG Disclosure and Firm Value

After a short introduction to why the NFRD constitutes an ideal research setting for a paper on the association of ESG disclosure and firm value, an overview of the scientific evidence on mandatory ESG information disclosure and its effect on firm value is laid out. First, concerning non-European countries and second, concerning European countries. As mentioned, this work does not analyse the channels through which value is generated or destroyed by an ESG mandate and therefore, no literature review on studies analysing these drivers will be given. Moreover, following the classification into first- and second-order consequences, this review of the existing literature shall mainly focus on second-order NFRD effects pertaining to capital markets, especially firm value (measured as Tobin's Q).

As constituted by various scholars, the NFRD – due to its nature of being a discrete, external shock to governance in the EU causing an exogenous and random assignment of control and treatment groups – provides an optimal setting to improve the credibility of the potential association of firm value and the NFRD tested in this paper. In a voluntary setting, researching ESG disclosure falls prey to a dual-selection problem as it depends on voluntary ESG activities of companies and their choices in reporting about these (Christensen

[5] Carrots & Sticks is a joint project between KPMG, the GRI, UNEP, and the Centre for Corporate Governance in Africa from the University of Stellenbosch. It assesses the regulatory landscape of non-financial and sustainability reporting (Gibbons, 2020; Van der Lugt et al., 2020).

et al., 2021; Fiechter et al., 2022; Ioannou & Serafeim, 2017; Ottenstein et al., 2022). As Atanasov and Black (2016) point out, so-called "legal shocks" come from legal rules or changes in those – like the NFRD – and are the most convincing for research designs.

Overall, the literature on the disclosure of ESG information differs between voluntary and mandatory disclosure. Within the literature on mandatory disclosure, a further differentiation can be drawn between first- and second-order consequences. First-order or direct consequences refer to effects like reporting quantity or quality attributable to a change in the regulatory environment. Second-order or indirect consequences include effects on the firm itself, investors and lenders, or society at large (Gulenko, 2018; Ottenstein et al., 2022). It is perhaps striking that even though the EU emphasizes its high expectations on second-order consequences in a memo released around the time of the mandate of the NFRD, most scientific publications have been conducted in the field of voluntary disclosure at that time. Only after the construction of the NFRD more publications on mandatory disclosure were released (Christensen et al., 2018; European Commission, 2014b; Gulenko, 2018).

Empirical evidence agrees that voluntary and mandatory ESG disclosure affects firm value. However, current literature does not clearly answer whether this effect is positive or negative[6]. Not only is literature regarding firm value implications of ESG disclosure mandates inconsistent in its findings, but it also is relatively scant and more recent since this type of disclosure became more common relatively recently compared to voluntary reporting (Christensen et al., 2021; Hummel & Jobst, 2022; Mittelbach-Hörmanseder et al., 2021).

A literature review dealing with the firm value implications of ESG disclosure mandates in countries outside the EU also shows mixed results. Four of the five mentioned relatively recent publications report a positive and only one a negative association of mandated ESG reporting with firm value.[7]

In the research for this work, one of the most cited contributions was produced by Ioannis Ioannou and George Serafeim in 2017. Their study explores the implications of ESG disclosure regulations in China, Denmark, Malaysia, and South Africa from just before 2011 utilizing a propensity score matched, two-staged differences-in-differences and IV setting. They find that treated companies significantly augmented ESG disclosure after the introduction of a disclosure mandate. This resulting increase, they conclude, is associated with increases in firm value, measured by Tobin's Q. Ioannou and Serafeim point out that it is challenging to separate the reporting effects from the effects of possible changes in the underlying ESG or other firm activities around the directive (Ioannou & Serafeim, 2017).

Approaching worldwide ESG disclosure mandates from a Mergers & Acquisitions (M&A) angle, T. Li et al. (2023) employ a difference-in-difference approach with a global M&A and divestiture data set spanning the years 2000 to 2020. They find that acquisitions following mandated ESG disclosure increase companies' ESG performance and may create long-term firm value. Using Tobin's Q as a proxy for firm value, their findings indicate a positive and statistically significant impact of acquisitions following ESG disclosure mandates on firm value after the adoption year of the respective mandate. In contrast, T. Li et al. (2023) state that when weak-ESG assets are divested following a mandate, divesting firms only experience a short-term improvement in their ESG scores, which does not translate into a higher firm valuation (T. Li et al., 2023).

In a similar vein, two studies examine Integrated Reporting in South Africa – a subset country of Ioannou and Serafeim's work – and find a positive relationship between mandated ESG disclosure and firm value. Lee and Yeo (2016) further conclude that following the mandated disclosure of Integrated Reporting through the King III Report on Corporate Governance in 2010, results of an on average positive relationship suggest that the benefits of the mandate exceed its costs. In a similar study, Barth et al. (2017) analyse – among other things covered – how firm value (measured by Tobin's Q) is positively associated with the quality of companies' Integrated Reporting after its introduction. Barth et al. (2017) judge their work's findings to be consistent with the results drawn out by Lee and Yeo (2016).

Moving from the African to the Asian continent, Chen et al. (2018) exploit an ESG disclosure mandate introduced on certain stock exchanges in China in 2008. Using a difference-in-difference design with a sample of 3,120 firm-year observations, they find that mandatory ESG disclosure is negatively associated with firm value and investment. Chen et al. conclude that their findings indicate that mandatory CSR disclosure changes firm behaviour and produces positive externalities, thus leading to costs for shareholders (Chen et al., 2018).

Similar results are reported by Manchiraju and Rajgopal (2017), who study Clause 135 of the Companies Act introduced in India in 2013, requiring companies above a certain threshold to spend 2% of their average net profits of the prior three years on ESG activities. Like the NFRD, Clause 135 is constructed in a "comply-or-explain" approach. Importantly, this law and the study concern the value implications of ESG *activities* and follow a mandate on ESG spending and *not* reporting. Utilizing an event study approach combined with a regression discontinuity design, Manchiraju and Rajgopal calculate a 4.1% drop in the stock price of companies subject to the mandate after the introduction of the legislation. Interestingly, they state that the mandate does not negatively affect companies that spend more on advertising. Measured by Tobin's Q, the long-term firm value is negatively influenced by the ESG mandate. The same mandate is exploited by Bha-

[6] Despite the mixed literature, most of the empirical evidence imply a positive association between voluntary disclosure of ESG information and firm value. This indication needs to be treated, as mentioned, with carefulness due to the dual-selection problem (Christensen et al., 2021; Hummel & Jobst, 2022; Mittelbach-Hörmanseder et al., 2021).

[7] The three papers concerning India are excluded as Clause 135 targets ESG spending instead of reporting.

gawan and Mukhopadhyay (2019), who apply a difference-in-difference setting combined with matching to evaluate the impact of Clause 135 on Tobin's Q as a measure of firm value. Contrasting the findings of Manchiraju and Rajgopal, Bhagawan and Mukhopadhyay find a positive and statistically significant impact on firm value. A third study harnessing Clause 135, published by Jadiyappa et al. (2021), results in a positive association of firm value (Tobin's Q) with increased ESG spending as well.

In the United Kingdom (UK), every firm was made to report comprehensive data on their GHG emissions in their annual reports by "The Companies Act 2006 (Strategic Report and Directors' Report) Regulations 2013", passed in July 2013. Krueger (2015a) leverages this mandate to investigate the effect of mandated greenhouse gas emissions disclosure on firm value (measured by Tobin's Q). Krueger not only uncovers that companies more significantly affected by the regulation were more strongly positively affected in valuation but also finds that an increase in valuation was most significant for larger firms and firms from carbon-intensive industries such as oil and gas. Compared to other works that do not, Krueger analyses the channels through which carbon disclosure affected firm value and indicates that capital market effects have more influence on value than real effects. According to his research, the mandate, on average, increased market liquidity and lowered information asymmetries (Krueger, 2015a).

Focussing on literature concerning the NFRD in Europe, in contrast to the studies relating to non-European countries/worldwide settings, the three studies mentioned show negative results regarding the equity market perception of mandatory ESG disclosure. Next to the study of Ioannou and Serafeim (2017), Grewal et al. (2019) was one of the studies most cited in the literature consulted for this work. In their study, Grewal, Riedl, and Serafeim assessed the stock returns of affected companies around significant events leading up to the entry into force of the NFRD. They found evidence that the NFRD, on average, leads to a negative market reaction. This negative reaction was more strongly pronounced for firms with lower pre-directive ESG performance and -disclosure levels. Grewal et al. (2019) conclude that investors evaluate the NFRD to be, on average, costly to firms and especially to those with weak pre-directive ESG performance and disclosure levels, as they would be forced to disclose additional ESG information. According to the study, this negative reaction is linked to proprietary and political costs.

Fiechter et al. (2022) inspect real effects following the introduction of the NFRD and document that affected companies increased their ESG activities and that they started doing so before the NFRD came into force in 2017. These effects were more pronounced for companies with lower pre-directive levels of ESG reporting and activities (high exposure[8]). For the measures of total ESG activities and the social score measure, Fiechter et al. (2022) outline increases for high-exposure firms as early as 2014. The authors conclude that the treatment effects are economically meaningful, with an increase in ESG for the mentioned high-exposure firms of 1.3% (2014), 2.4% (2015), 3.8% (2016), 3.8% (2017), and 6.8% (2018), relative to the base year 2013. No significant treatment effect for the environmental score measure was discovered on average. The authors hypothesize three potential explanations for the pre-directive increase in ESG activity and reporting: Internal development and efficiency increases from preparations for the mandate, increased public attention around the NFRD (see also Christensen et al., 2017; Grewal et al., 2019), and anticipation of potential adverse stakeholder reactions following mandated disclosure (including anticipation of potential stricter regulation see also Leuz and Wysocki (2015) and Manchiraju and Rajgopal (2017)). Besides, in concert with the results of Grewal et al., Fiechter et al. state a negative effect of the directive on Tobin's Q for EU companies, which is more distinct for high-exposure firms than low-exposure firms. However, that difference is statistically insignificant. Finally, as Fiechter et al. point out, the negative effect of the NFRD on Tobin's Q is concentrated in high-exposure firms with comparatively lower increases in ESG activities following the NFRD (Fiechter et al., 2022).

In line with the findings of Fiechter et al. (2022), Mittelbach-Hörmanseder et al. (2021) indicate a negative association of share prices of STOXX Europe 600 noted companies with mandatory ESG reporting following the introduction of the NFRD. They use textual analysis with a self-constructed topic-specific disclosure measure for the period between 2008 and 2016, which indicates an either positive or statistically insignificant relationship between ESG disclosure and share prices for the pre-announcement period (Mittelbach-Hörmanseder et al., 2021). For example, Rossi and Harjoto (2020) report a positive correlation of Tobin's Q following two Italian Legislative Decrees 231/2001 and 254/2016. The problem with this work lies in, first, the small number of companies in the sample (N=156) and even more so second, in the entanglement of mandatory (Decree 254/2016) and voluntary (Decree 231/2001) disclosure which, as learned from extant literature, delivers divided results. Thus, this work will not consider Rossi and Harjoto's results.

In a nutshell, it seems ex-ante unclear how mandatory ESG disclosure may impact capital markets and, in particular, firm value based on the current theoretical discourse. Resultingly, this work aims to shed more light on the topic and add to the scant evidence literature offers. It has to be noted that the empirical analysis conducted in this work does not cover the channels of value creation (such as the increased level of disclosure or the changes in firm behaviour).

[8] According to Fiechter et al. (2022), exposure firms are firms with low levels of pre-directive ESG performance (activities) and disclosure (reporting). Fiechter et al. measure *reporting* with a self-constructed variable consisting of various ASSET4 categories. *Activities* are measured using the Social and Environmental score provided by ASSET4 (Fiechter et al., 2022).

3.2. Hypotheses Development

In light of the theoretical foundation and the current state of scientific literature, the need to further empirically evaluate the effect of the NFRD on firm value arises. First, the absolute effect of the NFRD on firm value for the population of subject firms is hypothesized. Afterwards, cross-sectional effects depending on specific firm characteristics potentially leading to varying investor reactions are predicted. In total, five central predictions are made.

Taking up the theoretical perspectives unfolded in Chapter 2.3 again, it became clear that mandatory disclosure regulation seeks to increase firms' ESG information transparency by enforcing the disclosure of both positive and negative information, which otherwise may not be released (Ioannou & Serafeim, 2017; Verrecchia, 2001). Following Christensen et al. (2021) and Fatemi et al. (2018), it can be concluded that the magnitude of the capital market reaction to the passage of the NFRD in 2014 – in form of an adaption of firm value – fundamentally relies, first, on the net benefit to the firm on an aggregate level (as valued by capital markets), and second, on the level at which companies withhold material ESG information prior to the announcement of the NFRD.

Literature regarding mandatory financial disclosure and voluntary ESG disclosure that finds a positive association with firm value typically suffers from the dual-selection problem, which is less prevalent in the mandatory context (Christensen et al., 2021; Hummel & Jobst, 2022). In contrast to the voluntary setting, the implications of mandatory ESG disclosure for firm value are ex-ante unclear. To hypothesize the aggregate effect of the mandated disclosure in the European Union, costs and benefits to the affected firms, pre-directive disclosure levels, market valuation of information, and flaws in the construction of the mandate need to be evaluated from a capital market perspective.

On the benefit side, literature suggests that more and better ESG information can benefit capital markets through a lowered return threshold at which investors may be willing to invest in a firm, lower information risk (resulting in e.g., improved performance prediction), investor investment efficiency, greater liquidity, and lower cost of capital (Amihud & Mendelson, 1986; Chang et al., 2022; Christensen et al., 2021; Constantinides, 1986; Easley & O'Hara, 2004; Grewal et al., 2019; Krueger et al., 2021; Lambert et al., 2007, 2012). Moreover, the more and better disclosure of ESG information can change firm behaviour or mitigate externalities. It can incentivize firms to adapt behaviour and thus become more efficient. Firms may ramp up their carbon emission reduction efforts, improve employee retention and engagement, increase efficiencies in their supply chain or buyer/supplier relationship or improve overall safety and quality measures (Darendeli et al., 2022; Ioannou & Serafeim, 2017; La Porta et al., 2000; Wang, 2023). For example, Konar and Cohen (1997) and Schlenker and Scorse (2017) found that a rise in information availability increases operation efficiency and environmental performance. Cutler et al. (2004) show evidence that this increased information availability can improve surgery results. As Eccles et al. (2014) and Ioannou and Serafeim (2017) point out, the positive effects can derive either from the regulation itself or since the market views the mandated disclosure regulation (here: the NFRD) as a strong signal of a government's commitment to sustainability, the role of ESG information for society, and potential (more severe) regulation in the future. Indeed, during the time of the introduction of the NFRD, the Paris Agreement, an international treaty on climate change, was adopted by 196 countries at the United Nations Climate Change Conference (COP21) in Paris on December 12, 2015. This landmark event showcased the severeness of climate change and government commitment to sustainability – and thus also ESG information disclosure – since, for the first time, a binding agreement unifies all nations to combat climate change and its effects (UNFCCC, n.d.). All of the benefits mentioned may be viewed positively by investors and thus increase firm value.

Nevertheless, research suggests that markets do not efficiently price ESG information and its benefits. Choi et al. (2020) suggest that investors underreact to climate change risks following the observation that stocks of carbon-intensive firms underperform firms with low carbon emissions in abnormally warm weather. Additionally, they discover that in contrast to institutional investors, retail investors sell carbon-intensive firms in such weather. In a similar vein, Stroebel and Wurgler (2021) and Krueger et al. (2020) find that a significant majority of financial economists judge financial markets to underestimate climate risk and that institutional investors believe that equity valuations do not entirely reflect climate risk. Combining these relatively recent insights with the steep increase in interest in sustainability and ESG information and their incorporation in investment decisions in the last years is outlined in Chapter 2.3 and the introduction, it becomes evident that the mispricing of ESG information must have been even more pronounced during the time of the NFRD passage. At that time, the standalone benefits resulting from the passage of the NFRD in 2014 may have been viewed as significantly less value-enhancing than they were actually.

Contrasting the benefits, investors may anticipate various sources of cost. As mentioned earlier, these can be split into direct and indirect costs. Direct costs resulting from a mandated CSR disclosure regime stem from complying with the NFRD through implementation efforts or operating costs. As Grewal et al. (2019) and De Groen et al. (2020) point out, these costs are relatively lower than others. More significant for the value relevance of the NFRD are political and proprietary costs. Together, these form the indirect cost part. Political costs relate to current and priced-in future costs stemming from the politically induced pursuit of an investment, which is perceived as an investment carrying a negative net present value to shareholders. These can include costs for expanding and adjusting ESG activities to enhance ESG performance (Christensen et al., 2021). Proprietary costs refer to a potential decrease in the competitiveness of a firm subject to the mandate and thus influence its current and future attractiveness for investors. A firm will moreover carry costs

if it decides to maintain its weak ESG performance and, e.g., pay for penalties or if it decides to improve its ESG performance (Chang et al., 2022; Grewal et al., 2019; Healy & Palepu, 2001; Jensen & Meckling, 1978; Watts & Zimmerman, 1978). Furthermore, the mandated ESG disclosure in the EU could have levelled the playing field and thus prevented firms with good ESG performance from setting themselves apart from poor-performing firms. To highlight this, good performers would have needed to increase their efforts, resulting in higher costs (Ioannou & Serafeim, 2017; Mittelbach-Hörmanseder et al., 2021).

Based on the assumption that mandatory ESG disclosure reduces information asymmetry, next to the net benefits to firms subject to the NFRD, a prediction for the aggregate effect of the NFRD on firm value necessitates an evaluation of the extent to which firms withheld material ESG information before the passage of the mandate. If, prior to the mandate, most firms affected had disclosed a non-financial report with material ESG information, few new information would enter the market and produce a negligible aggregate market reaction in firm value.

Since in a voluntary setting that existed up until the passage of the first mandatory disclosure setting in the EU in 2014, following the assumption that firms behave rationally and make optimal decisions, one can follow that all firms where net benefits existed prior to the regulation, these firms would have already published ESG information voluntarily. Since with the NFRD, firms that until then would have not reported would now need to report even under negative net benefits. The negative net benefits would be negatively valued by capital markets and imply a negative firm value association with the NFRD. The reduced dual-selection problem often brought up to explain the relationship between ESG disclosure and firm value in a voluntary setting could be a viable explanation (Christensen et al., 2018; Mittelbach-Hörmanseder et al., 2021).

As pointed out, firms that did not report until the passage of the NFRD are likely to have a negative net benefit from reporting ESG information since they assume that the reporting consequences will incur higher costs than benefits for the firm. Hence, if the firm reports ESG information for the first time under the NFRD, it is likely that more negative than positive ESG news will be released into the market. Evidence suggests that markets respond in the same direction as the news or in "quite a few cases" (Christensen et al., 2021, p. 1199) asymmetrically. Flammer (2013) and Crifo et al. (2015) suggest that investors put more weight on negative ESG news. Like Flammer and Crifo et al., Krueger (2015b) similarly concludes that investors respond negatively to bad ESG information and have a weakly negative reaction to good information.

A mechanism reducing the market reaction's magnitude in either direction, often pointed out to be weakening the NFRD, may be the comply-or-explain approach (see Chapter 2.2) that the NFRD uses. This leaves companies the chance to omit material information that is potentially value-destroying. In addition to the potentially weaker market reaction resulting from the comply-or-explain approach, the generally vague definition of what and how to report on ESG matters in the directive[9] leaves companies with more room for cherry-picking, greenwashing, and publishing irrelevant information. In addition, the information published under the NFRD must not be assured by an external auditor. The auditor merely has to assure whether a report has been published or not for a firm to comply with the directive (Agliati, 2021; Christensen et al., 2021; Hummel & Jobst, 2022; Mittelbach-Hörmanseder et al., 2021). The missing strictness in the construction of the directive will reduce the magnitude of the market reaction to the NFRD since it is to be assumed that less material and value-relevant information will end up being released.

Resulting from the preceding discussion on the costs and benefits to individual entities, in line with the rather negative value implications of the directive that literature points out, the extent to which firms withheld material ESG information prior to the passage of the mandate, the valuation of the newly released information, and the flaws in the construction of the mandate, there will likely be a weakly negative association between the NFRD and firm value on an aggregate level following the passage of the NFRD.

Hence, the first hypothesis is defined as:

> ***Hypothesis 1a (H1a):*** *On an aggregate level, the NFRD is likely to negatively affect firm value.*

Carnini Pulino et al. (2022), Fiechter et al. (2022), Grewal et al. (2019), Ioannou and Serafeim (2017), and Mittelbach-Hörmanseder et al. (2021), among others, demonstrate a differentiated effect of the NFRD on firm value depending on the topics that the NFRD affects. In accordance with prior literature, this topical relevance will be measured by the individual E, S, and G scores of the combined ESG score from Refinitiv's database to understand the driving pillars of the combined ESG score. Therefore, 1b is sub-hypothesized:

> ***Hypothesis 1b (H1b):*** *Firm value effects will likely emerge with a different magnitude depending on which of the three ESG score subcomponents E, S, and G are concerned.*

A second sub-hypothesis deals with the effect of time on the magnitude of the combined effect of the NFRD on firm value. In accordance with the findings of Fiechter et al. (2022), who show that the anticipation effects prior to the entry-into-force in 2017 kicked in immediately after the passage of the directive in 2014 and that they are more pronounced in the late post-period than in the early post-period

[9] See European Commission (2014a, Article 1(1)): The report must include information "*to the extent possible*" to understand a firm's performance and its impact (dual-materiality). Moreover, risks linked to a firm's operations should be mentioned "*where relevant and proportionate*".

for the ESG scores[10], we can assume that the same association will hold for the association with firm value. Hence, third hypothesis 1c is:

> ***Hypothesis 1c (H1c):*** *The firm value results from H1a are likely to be more significant in the late post-period (2018-2019) than in the early post-period (2014-2017).*

The hereafter following cross-sectional prediction is based on the premise that the above elucidated individual levels of costs and benefits, the level of ESG disclosure prior to the directive and the market valuation of this information are individual to every firm.

Cross-sectional hypothesis ***H2*** follows from the premise that there are differences inherent to firms, which result in varying effects of the NFRD on firm value. A difference especially relevant in the context of ESG and mandatory disclosures is the ESG performance prior to the directive and the affiliation with a specific sector. For example, Krueger (2015a) shows that some industries have a more substantial environmental impact than others, such as the Oil and Gas and the Basic Materials industries (as measured by greenhouse gas emissions and thus primarily affecting the "E" of the combined ESG score). These firms are under higher public pressure and tend to release more ESG information to legitimize their actions, as research suggests (Cahan et al., 2016; Cho, Michelon, et al., 2015; Patten, 1992). Since this ESG information differs in relevance to market participants depending on the sector, the NFRD also has a different relevance (Eccles et al., 2012). Since sector affiliation tends to be associated with lower or higher ESG scores, this cross-sectional hypothesis is checked before the sectoral analysis for more robust results. It can be expected that sector affiliation is, on average, highly value relevant following the passage of the NFRD (Grewal et al., 2019; Ioannou & Serafeim, 2017; Jackson et al., 2020; Krueger, 2015a; Mittelbach-Hörmanseder et al., 2021). Accordingly, the value-relevance for pre-directive ESG performance and sectors can be predicted in the form of the following two cross-sectional hypotheses:

> ***Hypothesis 2a (H2a):*** *For firms with higher pre-directive ESG performance, the effect of the NFRD on firm value is less pronounced than for firms with lower pre-directive ESG performance.*

> ***Hypothesis 2b (H2b):*** *For some industries, the effect of the NFRD on firm value is more pronounced than for others.*

4. Sample, Summary Statistics, and Main Variables

4.1. Identification Strategy

This chapter opens with the strategy on how the main effects needed to answer the research question are extracted and how the time frame for the analysis is chosen. It continues with explanations of the sample construction and the matching procedure. In the perpetuation, summary statistics are reported for the matched control and treatment group. This chapter closes with elucidations on the two main variables in this work, ESG score and Tobin's Q.

In order to examine the effect of the NFRD on firm value, the passage of the directive in 2014 is used as a shock to ESG performance. The supranational and regulation-driven nature of the introduction across the EU, which transitioned ESG reporting from voluntary to mandatory, presents an ideal setting to exploit the effect of mandatory ESG disclosure on firm value. (Atanasov & Black, 2016; Christensen et al., 2017; La Torre et al., 2018; Leuz & Wysocki, 2015; Mittelbach-Hörmanseder et al., 2021).

Based on the remarks in Chapter 2.3 concerning the theoretical principles of ESG disclosure, an argumentative path for the examination of the research question must entail a reduction (increase) of frictions (benefits) precipitated by the disclosure mandate (e.g., reduced information asymmetry). It has to be noted here that this work does not aspire to analyse through which channel the market reaction to ESG disclosure (such as firm-specific and internalized market-wide cost or benefit assessment) shapes firm value but instead focuses on the aggregate effect of the NFRD on firm value itself.[11] Consequently, for example, firm value can be affected through reduced information asymmetry. Nonetheless, this supposition may not hold if firms do not increase their information disclosure and performance (measured by the ESG score (Fatemi et al., 2018)) due to, e.g., construction flaws inherent to the directive. Hence, it is per se not clear whether and how the passage of the NFRD influences firm value.

To answer the research question, in the first stage of the analysis, the legal shock will be tested for validity by estimating the direct effect of the NFRD on ESG performance (measured by the combined ESG score) of treatment firms (EU) compared to control firms (U.S.). This first stage in the 2SLS instrumental variable examination is conducted with a DID-analysis. Consistent with previous approaches (Allman & Won, 2022; Cahan et al., 2016; Fatemi et al., 2018; Fiechter et al., 2022; Gibbons, 2020; Ioannou & Serafeim, 2017) to show the impact of ESG disclosure or performance on firm value this serves as a falsification test that would eventuate in an invalid relationship if the result amounted to zero. Successively, instrumental variable regression is employed in the spirit of Fatemi et al. (2018) and Ioannou and Serafeim (2017) to estimate the effect of ESG performance on firm value.

4.2. Temporal Considerations

Since this study analyses the impact of the NFRD on firm value over time, relevant data for the sample at hand are retrieved over the period from 2011 to 2019. Taking into

[10] 1.3% (2014), 2.4% (2015), 3.8% (2016), 3.8% (2017), and 6.8% (2018), relative to the base year 2013 (Fiechter et al., 2022).

[11] For a deeper understanding of those channels, see e.g., Chang et al. (2022).

account the elucidations on the introduction period of the NFRD in 2013/2014 in Chapter 2.2, and the anticipatory effects prior to the entry into force in 2017, outlined by Cuomo et al. (2022), Fiechter et al. (2022), Grewal et al. (2019), and Mittelbach-Hörmanseder et al. (2021)[12], year 2014 is denoted as the shock date, in this research setting. Since all countries in the EU experienced the shock simultaneously, the differently timed transposition into national law in each country can be neglected in this research setting. Hence, the pre-period is defined as 2011-2013 and the post-period as 2014-2019.

The starting year of the sample is set to the year 2011 to minimize the impact of confounding events such as the introduction of the Dodd-Frank Act (2010) and the aftermath of the Global Financial Crisis (2008-2009) (IMF, 2013). The effects of the European Debt crisis, which lasted from 2009 until nearly the end of the 2010s, cannot be entirely excluded but are minimized since my sample starts after the crisis started and is present throughout the whole sample period (Frieden & Walter, 2017; Gross & Zahner, 2021). The end date for the sample is set to 2019 to first exclude the confounding effects of the COVID-19 pandemic that started to affect the economy in countries included in my sample at the beginning of 2020. Second, two ESG-disclosure-related events happened at the end of 2019. The adoption of the SFDR (December 2019) and the announcement of the Green Deal (December 2019) (European Commission, 2019b, 2019c). The temporal considerations result in the dummy variable, $Mandate_t$ that equals 1 for the years 2014-2019 and 0 for the years 2011-2013.

4.3. Sample Selection and Data

To test the directive's impact on firm value, firms exposed to the NFRD (treatment group) are compared to firms not exposed to the NFRD (control group). Given the condition that the directive pertains to all large undertakings in the EU, the analytical setup is not up to the standards of an experiment with a randomly assigned treatment group. Therefore, in the empirical analysis, EU public interest entities affected by the NFRD are compared to a propensity score-matched sample consisting of U.S. firms. The basis for the sample is a list of firms (primary equity securities) headquartered in the 28 EU countries and the United States (European Commission, 2014a). In line with Ottenstein et al. (2022), the sample still includes the United Kingdom, which left the EU on January 31, 2020.

The United States present a particularly suitable country to construct a control group since until now, and unlike many other countries in the world, the U.S. did not adopt any market-wide ESG-related disclosure regulations, afar from financial, risk, and litigation disclosure requirements concerning public companies with environmental issues, as well as human capital management related aspects from 2020 onwards (Christensen et al., 2021; Cifrino, 2023; Ioannou & Serafeim, 2017). In related works and during the construction of this sample, it became clear that Refinitiv's ESG database imposes a significant constraint on the sample size. The U.S. offers one of the most comprehensive country coverages of ESG data, easing control firms' matching (Fiechter et al., 2022; Refinitiv, 2022).

Contrary to Allman and Won (2022), who compare U.S. firms exposed to the NFRD to U.S. firms not exposed to the NFRD, a more significant number of scholars[13] do not use a sample of non-EU-reporting firms because there are only an infinitesimally small number of firms that are not in the EU that are obliged to report under the NFRD. In the "Study on the Non-Financial Reporting Directive – Final Report" published by the European Commission in November 2020, the authors explain that there are only 54 companies with shares and bonds listed on EU-regulated markets but domiciled outside the EU that are required to report under the NFRD. Most were in the UK (De Groen et al., 2020).[14] Only the introduction of the Corporate Sustainability Reporting Directive (CSRD; publication on December 16, 2022) required third-country issuers within the EU to prepare reports according to the CSRD (Baumüller & Grbenic, 2021; European Commission, 2022). Around the publication of the CSRD, many consultancies and advisors released information on the CSRD and its potential effects on their clients. Announcements from, e.g., law firms Gibson Dunn, White&Case, and Latham&Watkins or consultancy Deloitte show further proof that the NFRD virtually mandates no non-EU firm to report ESG information (Davies et al., 2023; Deloitte Touche Tohmatsu Limited, 2023; McGarry et al., 2022; Sagayam et al., 2022). Moreover, the European Parliament posted a press release on November 11, 2022, stating that the CSRD is extending the scope to "Non-EU companies with substantial activity in the EU" and "addressing "shortcomings in existing legislation on the disclosure of non-financial information [NFRD] perceived as largely insufficient and unreliable." (European Parliament, 2022, both p. 1). Following the clarification in this controversy, the control group used in this work is constructed using U.S. firms. This defines the start of the sample selection process with in total 23,383 firms (Table 1).

As a first step, firms operating in the Financials and Utilities sector (TRBC economic sector classification) are excluded as they follow different market mechanisms and are subject to distinctive regulations (Allman & Won, 2022;

[12] Krueger (2015a) uses the same anticipatory approach in a different DiD-Setting, analysing the effect of a mandatory greenhouse gas emissions disclosure law on corporate value, announced in 2012 in the United Kingdom.

[13] Scholars that do not classify non-EU firms as reporting firms in the context of the NFRD are e.g., Cuomo et al. (2022) and Fiechter et al. (2022), and Cicchiello et al. (2023) who compare EU and U.S. firms, Ottenstein et al. (2022) who compare EU with OECD less EU firms, and Z. Li and Jia (2022) who focus on a comparison of Danish, French, Norwegian, and Swedish firms with U.S. firms.

[14] Since the report was written *after* the United Kingdom left the EU, and hence UK firms are still included in the sample used for this work, the neglectable, small number of U.S. firms left in the group of these 54 companies will not be excluded from the sample.

Chen et al., 2018; Cuomo et al., 2022; Fama & French, 1992; Fuente et al., 2022; Ioannou & Serafeim, 2017). The inclusion in the TRBC sector categorization significantly reduces the sample size, as not all firms are covered in this classification. Nevertheless, in Refinitiv Eikon, this sector classification leaves the most extensive sample. After this stage, the sample consists of 13,543 firms. For each of these respective firms, the Refinitiv Eikon Excel-Add-In was used to obtain firm-level data concerning the EU thresholds of revenue, assets, and revenue.

In the next step, the sample is loaded into the Python code used for data cleaning, preparation, matching, and, ultimately, the DID analysis. It is narrowed down to the firms that are – and in the case of the U.S., firms would have been – affected by the NFRD by applying the restrictions established in Article (1) of the NFRD in combination with references to the Accounting Directive 2013/34/EU, Article (2). According to these, the NFRD only affects large undertakings that are public-interest entities with more than 500 employees on average. "Large undertakings[15]" exceed at least two of the three criteria, total assets of EUR€ 20,000,000, revenue of EUR€ 40,000,000, or number of employees more than 250 on average (European Commission, 2013). With the 500 employees pre-determined by the NFRD, this condition of the large undertaking is fulfilled, and only one of the two monetary size thresholds needs to be satisfied. To address potential data collection issues in the data basis provided by Refinitiv, a 5% safety margin is added on top of the respective NFRD thresholds, further reducing the sample size. Furthermore, firm years where all data points are missing are not imputed but excluded. As a result, 2,879 firms are identified in the sample.

Opposing previous literature (Allman & Won, 2022), it must be highlighted that firms below the thresholds set out above are deliberately not included in the control group because it is likely that ESG and firm value are moderated by firm size. Evidence is brought forward by D'Amato and Falivena (2020), who conducted an analysis based on a dataset of Western European listed firms and found that ESG and firm value are moderated by firm size and age with a negative association when small and/or new companies are studied.

To avoid potential country bias in my sample, firms that have been subject to the NFRD following a deviation from the EU definition when transposed into national law are not included in the sample. As mentioned in Chapter 2.2, 13 states have introduced a broader definition of the minimum threshold, ramping up the number of mandated companies to around 11,500 (De Groen et al., 2020). This divergence could be exploited in future research to further explore the effect of the NFRD on firm value.

The Refinitiv Eikon Data API is accessed and used via Python code to obtain all ESG and firm-level financial data.[16] After checking if at least one ESG score observation is available throughout the period from 2011 to 2019, firms that do not fulfil this requirement are dropped. This yields 2,231 firms. Moreover, firm years with missing variables are dropped. Firms with any missing values for all variables in the matching year 2013 are dropped. Refinitiv's coverage for the ESG score imposes, by the absolute number of firms the sample is reduced by, a significant constraint to the sample composition. Lastly, with these exclusions, the sample consists of 1,034 U.S. and EU firms before matching.

Matching

After identifying the sample of EU and U.S. firms and cleaning the data for the matching year, propensity score matching (PSM) is applied to account for fundamental differences between treatment and control firms. For all of these 1,034 firms, the complete data set for the observed period from 2011-2019 is drawn from the master data file. Manual random tests are conducted for firms' financial and ESG data validity. For this data set, no exclusion nor imputation of missing values is conducted as introducing bias to the sample can thus be prevented (Seaman et al., 2013). In my final data set after matching, missing values are within reasonable limits below the 25% threshold suggested by researchers such as Collins et al. (2001) or Graham (2003). This approach allows for a steadier view of firms' financial data as, in some cases, in the data of the sample used, ESG scores are missing where financial data are available. These missing values explain the differences in the number of observations in the tables and experiments conducted in the work at hand (Fiechter et al., 2022). Table 1 in section 4.4 *Summary Statistics* reports on the different variables in the sample.

Specifically, the matching is performed one year prior to the shock (matching in 2013) to account for the pre-directive anticipation effects. Matching prior to the shock reduces the possibility that the matching variables are affected by the treatment (Atanasov & Black, 2016; Fiechter et al., 2022; Flammer, 2015; Ioannou & Serafeim, 2017; Z. Li & Jia, 2022; Shipman et al., 2017). To obtain the control group, the sample is matched on firm-level characteristics that may covariate ESG performance and Tobin's Q (Bajic & Yurtoglu, 2018). These include ln($SIZE$) (natural logarithm of total assets), $SECTOR$ (TRBC sector classification), LEV (Leverage; total liabilities over total assets), ESG (ESG score), and Q_i (Tobin's Q, which is the market expectation about potential growth opportunities). The matching variables were selected after an analysis of matching variables from similar studies applied by Chen et al. (2018), Cicchiello et al. (2023), Cuomo et al. (2022), Flammer (2015), Gibbons (2020), Grewal et al. (2019), Ioannou and Serafeim (2017), Krueger (2015a), Z. Li and Jia (2022), and Ottenstein et al. (2022). Chapter 5.1 *Baseline Analysis* further expands on their relevance. All variables are explained in *Appendix A*. The matching is run using the Python PsmPy-package, developed by Kline and Luo (2022). It is "based on a logistic regression logit score where a match is selected using k-nearest neigh-

[15] Defined by directive 2013/34/EU.

[16] See Chapter 4.5 for a detailed explanation on why Refinitiv was used to obtain ESG data.

bours." (Kline & Luo, 2022, Abstract). Following Shipman et al.'s (2017) remarks related to PSM design choices and an exemplary application by Fiechter et al. (2022) in their assessment of the NFRD, the PSM is run without replacement and a caliper distance of 0.05 to increase the likelihood of high-quality matches and improve covariate balance without a too significant reduction of the sample size.[17] No winsorizing was conducted as all variables were checked for outliers manually, and only one outlier company was found and excluded. Closest neighbours for each firm are matched in the year 2013 based on Ioannou and Serafeim (2017):

$$Treatment_i = \tau_0 + \tau_1 \ln(SIZE)_i + \tau_2 SECTOR_i + \tau_3 LEV_i + \tau_4 ESG_i + \tau_4 Q_i + \varepsilon_i \quad (1)$$

Where i represents firms and τ_0 denotes the intercept of the equation. ε is the unobserved error term. The treatment categorization results in the dummy variable $Treatment_i$ that equals 1 if firm i is subject to the NFRD and 0 otherwise. The matching further reduces the sample size due to the matching without replacement. From in total 1,034 firms before matching, 708 are left in the sample after matching. The final sample includes 354 firms in the treatment and 354 firms in the control group.

4.4. Summary Statistics

Table 1 Panel A describes the selection approach for the sample used in this work. Firm observations span from 2011 to 2019 and from the 23,383 firms headquartered in the 28 EU countries and the U.S., 708 firms make it into the final sample.

Overall, out of the 28 EU countries included in the sample at the start of the selection process, only 17 made it into the final sample (Cuomo et al. (2022) also obtained a sample with firms based in 17 EU countries in a related study). Furthermore, as panel D of Table 1 depicts, 60% of firms in the EU treatment group are concentrated in the United Kingdom, France, and Germany. This distribution may be attributed to the coverage in the Refinitiv databases or the notation in major stock indices[18] (Ottenstein et al., 2022). Apart from these observations, the sample composition appears plausible as the United Kingdom, France, and Germany were the largest economies in the European Union in 2013/2014 in terms of GDP (World Bank, 2023).

A similar 60% weight can be observed in the sector distribution in the sample. For both control and treatment groups, the Consumer Cyclicals, Industrials, and Technology sectors combined comprise around 60% of the sample. In the control and treatment group, the best-performing sectors in the combined ESG score are Real Estate, Consumer Non-Cyclicals, and Healthcare. Worst performing sectors are different for both groups. The treatment group includes the Industrials, Basic Materials, and Energy sectors. In contrast, in the control group, the Industrials, Basic Materials, and Consumer Cyclicals sectors are the bottom three performers (Academic & Educational Services is excluded as it includes one non-representative observation). A more detailed view of the individual properties of both the control and treatment group is given in section 6.1 Descriptive Statistics.

4.5. Main Variables

Measuring Firm Value

The primary dependent variable used to measure firm value and assess the effect of the NFRD on the latter is Tobin's *Q*. This approach aligns with prior literature that commonly uses Tobin's Q as a proxy for firm value[19] (De Groen et al., 2020). In this work, Tobin's Q is calculated as the sum of assets and market value of equity minus common book equity (book value of liabilities) over the total assets (Ahmad et al., 2023; Barth et al., 2017; Daske et al., 2008; Ioannou & Serafeim, 2017; Manchiraju & Rajgopal, 2017; Pestana Pavan, 2020). Tobin's Q integrates an investor's evaluation of a firm's potential cash flows as well as its level of risk. Based on the theoretical perspectives regarding the reduction of information asymmetries and the effects on risk and cost of capital/cash flows outlined in Chapter 2.3, one can assume that ESG information and their disclosure have implications for firm value (Cahan et al., 2016; Easley & O'Hara, 2004; Grewal et al., 2019; Krueger et al., 2021; Lambert et al., 2012). Cahan et al. (2016) exemplify that the ESG information disclosure induced by the directive could hypothetically uncover sustainability-related market opportunities or good relationships with specific interest groups that are important for some investors, increasing the market's expectations of future cash flows and reducing investment risk and thus affect Tobin's Q. Adversely, when disclosed ESG information such as environmental management procedures or certain sustainability or humanitarian projects are contrasting market expectations of these, investors may reduce expectations for cash flows or increase expectations for firm risk, hence decreasing Tobin's Q.

Measuring ESG Performance

As pointed out previously, research commonly agrees that ESG performance is relevant to investors (Christensen et al., 2021; Eccles et al., 2011). One of the most common ways for investors to quickly assess a company's ESG performance is to look at its ESG ratings (Eccles et al., 2011; Fiechter et al., 2022; Ioannou & Serafeim, 2017). Mainly in the form of a cumulative number that allows investors to easily compare investment opportunities via a single number, which saves time and cost. In a comparison with numbers obtained from data

[17] To test sensitivity and ensure high-quality matching, the code was run with larger and smaller calipers or with replacement.
[18] FTSE100, DAX30, or the CAC40.
[19] Or market reaction (Agostini et al., 2022; Buallay, 2019; Conway, 2019).

Table 1: Sample Description

Panel A: Sample Selection

Selection Criteria	EU Sample	U.S. Sample	$\sum$	Unit
Start: Firms headquartered in EU28 and U.S.	**10,364**	**13,019**	**23,383**	Firms
Less observations of firms:				
Without TRBC Economic Sector classification	2,268	3,672	5,940	Firms
Without duplicates and TRBS Sector Utilities and Financials	1,539	2,361	3,900	Firms
Firms before NFRD filter	**6,557**	**6,986**	**13,543**	Firms
Without number of employees <525 AND <EUR€ 21,000,000 total assets or EUR€ 42,000,000 revenue	5,056	5,608	10,664	Firms
Firms before data cleaning	**1,501**	**1,378**	**2,879**	Firms
Without *ANY* ESG score value between 2011-2019	589	59	648	Firms
Without firms with rows with missing values for variables	14	12	26	Firms
Without firms with missing control variable data in 2013	409	762	1,171	Firms
Final sample before matching	**489**	**545**	**1,034**	Firms
Final sample after matching	**354**	**354**	**708**	Firms

Panel B: Sample Distribution per Sector - EU Firms

Sector	N	Firm Years	Firm years (%)	ESG Score	Env. Score	Soc. Score	Gov. Score	ln(SIZE)	Q	Lev
Consumer Cyclicals	81	712	22.682%	55.109	51.343	59.368	50.820	31.801	2.259	0.589
Industrials	79	706	22.491%	50.927	49.697	54.059	47.501	32.049	1.663	0.680
Technology	51	451	14.368%	54.066	48.318	57.437	53.865	32.179	1.950	0.603
Basic Materials	42	370	11.787%	53.312	50.506	54.427	56.486	32.120	1.544	0.541
Consumer Non-Cyclicals	41	365	11.628%	58.786	60.571	60.928	53.462	32.691	1.908	0.617
Healthcare	32	286	9.111%	55.616	48.617	62.014	51.399	32.053	2.852	0.529
Energy	19	169	5.384%	53.045	52.955	55.664	49.003	32.649	1.123	0.546
Real Estate	9	80	2.549%	60.855	70.391	62.762	47.300	32.760	1.141	0.553

Panel C: Sample Distribution per Sector - U.S. Firms

Sector	N	Firm Years	Firm years (%)	ESG Score	Env. Score	Soc. Score	Gov. Score	ln(SIZE)	Q	Lev
Industrials	84	749	23.725%	48.241	39.890	48.698	54.795	32.212	2.377	0.640
Consumer Cydicals	78	694	21.983%	53.422	45.018	56.835	53.381	32.364	2.454	0.673
Technology	54	482	15.268%	57.898	51.017	61.229	58.636	32.726	2.823	0.557
Basic Materials	41	368	11.657%	50.414	46.755	47.692	60.694	32.205	1.774	0.649
Consumer Non-Cyclicals	36	324	10.263%	64.490	61.634	65.532	65.604	33.417	2.605	0.697
Healthcare	26	228	7.222%	62.528	54.635	64.585	63.266	33.889	2.524	0.595
Energy	22	195	6.177%	56.910	54.410	55.454	63.843	34.125	1.301	0.521
Real Estate	12	108	3.421%	62.405	57.539	67.202	63.337	33.067	1.839	0.616
Academic & Edu. Services	1	9	0.285%	27.752	0.000	35.362	43.986	29.111	1.319	0.788

The shown table depicts the sample and its properties after filtering, data cleaning, and propensity score matching. All data is retrieved via Refinitiv Eikon Data API. PSM is conducted based on the pre-directive values of the matching variables in 2013 and uses a caliper of 0.05 without replacement. Variables used are: ln(SIZE), Sector association, Leverage, ESG score, and Tobin's Q.

Table 1 — continued

Panel D: Sample Distribution with Mean Values per Country

Couuntry	N	Firm Years	Firm years (%)	ESG Score	Env. Score	Soc. Score	Gov. Score	ln(SIZE)	Q	Lev
U.S.	366	3294	100%	54.823	47.926	56.177	58.271	32.688	2.366	0.625
UK	121	1089	33.060%	52.193	46.822	53.656	55.201	31.576	3.598	0.574
France	50	450	13.661%	60.666	67.949	65.674	46.626	33.333	1.710	0.628
Germany	45	405	12.295%	55.404	51.633	59.317	51.308	32.794	1.769	0.620
Sweden	20	180	5.464%	61.224	56.500	67.645	55.314	31.947	1.627	0.596
Ireland (Rep.)	19	171	5.191%	46.117	36.958	46.950	54.481	32.043	2.363	0.630
Netherlands	19	171	5.191%	59.967	58.385	65.426	54.044	32.943	1.584	0.640
Denmark	16	144	4.372%	53.695	49.560	58.012	49.004	31.102	3.696	0.503
Spain	16	144	4.372%	61.462	61.706	70.446	48.345	32.347	1.896	0.657
Belgium	11	99	3.005%	48.747	48.202	49.311	50.348	32.222	1.401	0.611
Finland	10	90	2.732%	56.109	62.233	60.497	42.748	31.094	2.216	0.580
Italy	10	90	2.732%	54.078	45.385	57.945	53.600	32.747	1.321	0.711
Poland	8	72	2.186%	38.419	32.704	34.034	47.537	31.353	1.186	0.568
Greece	7	63	1.913%	37.971	33.084	40.440	38.940	31.190	1.113	0.666
Luxembourg	5	45	1.366%	55.212	50.963	59.572	50.444	32.908	1.435	0.520
Austria	4	36	1.093%	49.905	51.859	50.598	46.803	32.522	1.259	0.622
Portugal	3	27	0.820%	58.574	61.911	63.211	45.550	32.322	1.791	0.700
Hungary	2	18	0.546%	50.362	46.914	53.741	44.491	32.453	1.232	0.419
$\sum$ // M EU firms	366	3294	100%	52.947	50.751	56.263	49.105	32.170	1.835	0.603

provider Bloomberg, Eccles et al. (2011) quantify the relevance of this combined score that summarizes a company's ESG performance: In a time frame of six months in 2010 and 2011, the data point "ESG Disclosure Score" was 60% more accessed than the second next data point, "GHG Scope 1". Following these capital market insights, Fatemi et al. (2018) built a model arguing that the value of a firm can be affected by ESG activities and the disclosure of those. To assess an association between the NFRD – which mandated the disclosure – and firm value, a measure for ESG performance is needed.

Investigating ESG performance, not only investors but also research commonly relies on ESG scores (Fiechter et al., 2022; Flammer, 2021; Grewal et al., 2019; Ioannou & Serafeim, 2017; Jackson et al., 2020; T. Li et al., 2023; Z. Li & Jia, 2022; Liang & Renneboog, 2020; Mittelbach-Hörmanseder et al., 2021; Ottenstein et al., 2022). In line with the relevance for capital market participants and previous research as well as due to the fact that it offers one of the largest and most reliable sources for ESG performance information, the Refinitiv ESG database (formerly Thomson Reuters ASSET4), specifically the Eikon Data API is used to obtain ESG scores for sample companies in this work (Eccles et al., 2014; Fiechter et al., 2022; Jackson et al., 2020; Mittelbach-Hörmanseder et al., 2021; Ottenstein et al., 2022).

Refinitiv's ESG score (and its sub-scores, Environmental (E), Social (S), and Governance (G)) is a comprehensive measure of a company's ESG performance, commitment and effectiveness grounded on publicly disclosed information. The score is based on thoroughly collected, verifiable reported data in the public domain, such as company reports, company websites, and other sources, such as newspapers and nongovernmental organizations. This information is only included in the score when the switch from private to public is made mandatorily or voluntarily (or coincidentally) by the company. This enables the score to be used as a proxy for ESG performance that reflects changes in the disclosure switch from voluntary to mandatory disclosure.

The score summarizes 630 firm-level ESG measures, which, in a smaller subset, are grouped into ten categories that constitute the three pillar scores and the final ESG score[20]. The individual ESG pillar score is a relative sum of the category weights, which vary per sector for the environmental and social categories and remain the same across

[20] For a more detailed look at the construction of the score, see Refinitiv's „ESG scores methodology" under: https://www.refinitiv.com/content/dam/marketing/en_us/documents/methodology/refinitiv-esg-scores-methodology.pdf.

industries for governance. Afterwards, the pillar weights are normalised to percentages between 0 to 100. Refinitiv updates all its scores on a weekly basis (Refinitiv, 2022).

5. Research Design

5.1. Baseline Analysis

In order to identify the effect of the Non-Financial Reporting Directive on firm value, an instrumental variable model, specifically a 2SLS research setting in the sense of Fatemi et al. (2018) and Ioannou and Serafeim (2017), is applied.

The DID approach is commonly applied by scholarship to evaluate the effects of regulatory policies and in studies similar to the work in hand (Atanasov & Black, 2016; Bajic & Yurtoglu, 2018; Christensen et al., 2017; Daske et al., 2008; Fiechter et al., 2022; Flammer, 2015; Ioannou & Serafeim, 2017; Krueger et al., 2021; Ottenstein et al., 2022). The average treatment effect can be estimated by observing the change in the dependent variable for the treatment group compared to the control group both before and after the introduction of the policy studied. Hence, in this work, EU firms (treatment group) are compared to U.S. firms (control group) before and after the adoption of the NFRD in 2014. In absence of this directive, trend outcomes for both treated and control firms would remain in parallel trends. This functions as the central assumption for the DID approach used in this work.

The IV2SLS (Instrumental Variable Two Staged Least Squares model) was chosen on the basis that there is a well-researched connection between ESG performance and firm value (e.g., Bajic and Yurtoglu, 2018; Fatemi et al., 2018) and to monitor the value implications of ESG performance (measured through ESG scores) in the presence of mandatory ESG disclosure regulation. Furthermore, when examining the effect of the NFRD on firm value, potential endogeneity issues need to be addressed. This can be done with the IV approach (De Andrés et al., 2017; Fatemi et al., 2018; Fuente et al., 2022; Jiao, 2010). As Angrist and Krueger (2001) mention, "the most efficient way to combine multiple instruments is usually two-stage least squares" (p.70). The instruments chosen for the IV must fulfil two conditions. First, the instrument must be correlated with the endogenous variable (relevance condition) and second; the instrument must be uncorrelated to the error term (exogeneity condition) (Bascle, 2008).

Henceforth, following suggestions and applications by Angrist and Krueger (2001), Angrist and Pischke (2009), Fatemi et al. (2018), Huntington-Klein (2021), and Ioannou and Serafeim (2017), in the first stage, the endogenous variable ESG_{it} is predicted using a DID OLS. In the second stage, the predicted ESG score (ESG_{it}) from the first stage regression is inserted into the second-stage equation using the interaction term $Treatment_i * Mandate_t$ as an instrument. Thus, the IV model can yield results for the impact of the NFRD on firm value through ESG performance.

Confirming the Shock to ESG Performance – First Stage of IV

Prior to the second stage, which tests the effect of the directive on firm value through ESG performance, the legal shock will be tested for validity and significance by exhibiting that the ESG performance (measured by the combined ESG score) of treatment firms significantly increases after the passage of the NFRD in 2014 compared to control firms. With the DID design, the treatment effect for the sample with all years between 2011 and 2019 is estimated. In the sense of Ioannou and Serafeim (2017) and Krueger (2015a), the following OLS regression is used for the first stage of the IV approach:

$$ESG_{it} = \beta_0 + \beta_1(Treatment_i * Mandate_t) + \beta_2 \ln(SIZE)_{it} + \beta_3 LEV_{it} + \alpha_i + \mu_t + \gamma_{jt} + \varepsilon_{it} \quad (2)$$

ESG_{it} denotes the ESG score from Refinitiv for firm i in year t. Drawn from equation (1), $Treatment$ takes the value of 1 if firm i is subject to the NFRD and 0 otherwise. The dummy variable $Mandate_t$ marks years where the directive was adopted and in force (2014-2019) with 1 and 0 otherwise (2011-2013).

With the endogenous nature of ESG disclosure, the analysed association between ESG performance and firm value may be driven by reverse causality or correlated or omitted and thus biased variables. Hence, following Angrist and Pischke (2009) and Gormley and Matsa (2012) and the exemplary application of their findings in Cahan et al. (2016), Cuomo et al. (2022), Ioannou and Serafeim (2017), Jackson et al. (2020), and Krueger (2015a), firm, year, and year-sector fixed effects, $(\alpha_i + \mu_t + \gamma_{jt})$ are incorporated in equation (2). These fixed effects increase the quality of the findings as they account for unobservable differences (e.g., legal system, political events, cultural influence, increasing influence of ESG) among sample firms. Among all equations for ***H1***, standard errors are clustered at the firm level to account for potential serial correlation in the error terms (Gibbons, 2020; Ioannou & Serafeim, 2017; Krueger, 2015a; M. Lang et al., 2012; Petersen, 2009).

The average treatment effect to confirm the shock induced by the passage of the NFRD in 2014 is depicted in the β_1 on the interaction term $Treatment_i * Mandate_t$, which displays the difference after the switch from voluntary to mandatory ESG disclosure for treated firms in comparison to the difference for the control group. A positive coefficient β_1 would reveal an increase in the average ESG score after the passage. To complete the first stage, which first assesses the effect of the NFRD on ESG performance and to mitigate the potential omitted variable bias, two control variables are added: $\ln(SIZE)$ (natural logarithm of total assets) and LEV (leverage; total liabilities over total assets). These firm characteristics are thought to be potentially linked with both ESG performance and Tobin's Q (Bajic & Yurtoglu, 2018).

These controls have been used in prior literature related to the work at hand (Cahan et al., 2016; Carnini Pulino et

al., 2022; Cicchiello et al., 2023; Daske et al., 2008; Dhaliwal et al., 2011; Eccles et al., 2014; Fiechter et al., 2022; Flammer, 2015; Ioannou & Serafeim, 2017; Jackson et al., 2020; Krueger, 2015a). Specifically, $\ln(SIZE)$ is used as a control since larger firms are inclined to release more ESG-related information and have higher ESG performance and related scores (Baran & King, 2014; Drempetic et al., 2020; M. Lang & Lundholm, 1993). LEV is accounted for because a monitoring role is inherent to debt servicing, and holders of debt request more release of information (Dhaliwal et al., 2011; Leftwich et al., 1981). Furthermore, leveraged firms are riskier, focused on short-term investments, have lower future growth[21], and tend less to reveal ESG information and hence negatively affect ESG ratings (Fama & French, 2002; Krueger et al., 2021; L. Lang et al., 1996).

In the sense of Angrist and Pischke (2009) and an exemplary application of their "good/bad control"-definition in a work examining the impact of a greenhouse gas emission disclosure mandate on firm value by Krueger (2015a), no further variables (e.g., ownership concentration, profitability, or research and development expenses) are included in the equations applied in this work.

Second-Stage of IV

In the second step of the IV, equation (3) tests the effect of the directive on firm value through the estimated ESG score from the first stage. With the dependent variable firm value, measured using Tobin's Q (Q_{it}), the following second-stage IV equation, similarly applied in Ioannou and Serafeim (2017), is estimated to answer ***H1a*** and slightly altered, ***H1b*** and ***H1c***:

$$Q_{it} = \vartheta_0 + \vartheta_1 \widehat{ESG}_{it} + \vartheta_2 \ln(SIZE)_{it} + \vartheta_3 LEV_{it} + \alpha_i + \mu_t + \gamma_{jt} + \varepsilon_{it} \quad (3)$$

In equation (3), $\widehat{ESG}_{it}$ serves as an independent variable with the hat denoting the predicted ESG score from Refinitiv for firm i in year t. ϑ_1 captures the effect of ESG scores on firm value. If ϑ_1 is positive and significant, it would imply that the directive positively affects Tobin's Q through the mandated disclosure of ESG information. The set of control variables employed in equation (3) – including the fixed effects – is the same as in equation (2) since these firm characteristics are thought to be potentially linked with both ESG performance and Tobin's Q (Bajic & Yurtoglu, 2018). Herewith, this approach expands scholarship that evaluates the impact of ESG performance on firm value as well (Fatemi et al., 2018; Fiechter et al., 2022; Fuente et al., 2022; Gibbons, 2020; Ioannou & Serafeim, 2017; Lee & Yeo, 2016; Manchiraju & Rajgopal, 2017). Contrary to ***H1a***, which assesses the impact of the NFRD on firm value on an aggregate level, ***H1b*** investigates the effects of the individual ESG score components E, S, and G on firm value to understand which components may be driving the aggregate effect on firm value. ***H1b*** will be tested in accordance with the procedure applied in Ioannou and Serafeim (2017) by running equation (2) and (3) with the subcomponent scores obtained from Refinitiv for E ($ESG_{E,it}$), S ($ESG_{S,it}$), and G ($ESG_{G,it}$), respectively.

To assess ***H1c***, whether the effects of the NFRD on firm value are more significant in the late post-period (2018-2019) than in the early post-period (2014-2017), a model estimating *yearly* treatment effects in relation to the base year 2013, developed by Fiechter et al. (2022) is deployed. A two-staged DID approach is estimated to utilise the predicted ESG score from the first stage (4; $\widehat{ESG}_{it}$) as a control variable in the second stage (5). The first- and second-stage results can be observed in the respective coefficient on the interaction term. Equations (4) and (5) are established to answer ***H1c***:

$$ESG_{it} = \beta_0 + \sum_n (\beta_n Year_n * Treatment_i) + \sum_q \beta_q \ln(SIZE)_{it} + \sum_p \beta_p LEV_{it} + \alpha_i + \mu_t + \gamma_{jt} + \varepsilon_{it} \quad (4)$$

$$Q_{it} = \vartheta_0 + \sum_n (\vartheta_n Year_n * Treatment_i) + \sum_q \vartheta_q \ln(SIZE)_{it} + \sum_p \vartheta_p LEV_{it} + \sum_r \vartheta_r \widehat{ESG}_{it} + \alpha_i + \mu_t + \gamma_{jt} + \varepsilon_{it} \quad (5)$$

Applying Larcker and Rusticus (2010) suggestion – which found application in Fatemi et al. (2018) – that results of a 2SLS approach should be compared to those obtained after running an OLS approach, the yearly effects estimated in equations (4) and (5) with an OLS regression simultaneously serve as an early robustness test to the findings of ***H1a*** which resulted from a 2SLS model.

5.2. Cross Sectional Analysis

To test whether the sample specification affects any of the results, cross-sectional analyses are performed on two sample specifications. As discussed in the hypothesis development section, the two cross-sectional predictions (***H2a*** and ***H2b***) are based on the premise that individual levels of costs and benefits, the level of ESG disclosure prior to the directive and the market valuation of this information are individual to every firm. A difference especially relevant in the context of ESG and mandatory disclosures is the affiliation with a specific sector and the pre-directive ESG performance, as prior research related to mandatory disclosure regulations shows that these attributes are entangled (Jackson et al., 2020; Krueger, 2015a; Mittelbach-Hörmanseder et al., 2021).

[21] According to L. Lang et al. (1996, Abstract), leverage is "negatively related to growth for firms whose growth opportunities are either not recognized by the capital markets or are not sufficiently valuable to overcome the effects of their debt overhang.".

To analyse the first cross-sectional hypothesis, ***H2a***, whether the value effect is less pronounced for firms with higher pre-directive ESG performance than for firms with lower pre-directive ESG performance, a partition "*ESG*" based on the ESG performance prior to the directive is introduced. Treatment firms with their respective controls are allocated to group "LowESG" if they have an ESG score in the lower tercile (lowest 33.3% of firms) in the matching year 2013. Similarly, the group "HighESG" comprises treatment firms in the upper tercile (66.6% and above) of ESG scores and their matched controls. This method is similar to the one used by Fiechter et al. (2022), Grewal et al. (2019), Ioannou and Serafeim (2017), and Jackson et al. (2020). Based on the anticipatory effects reported in, e.g., Fiechter et al. (2022) and the changes in firm behaviour already as early as 2014, each partition, LowESG and HighESG is run with their respective matched control firms and the same IV2SLS regression with equation (2) and (3) used to test ***H1a***. For the cross-sectional analyses, all three fixed effects stay in place, in line with Ioannou and Serafeim (2017). Standard errors for the partitions on ex-ante ESG performance are clustered at the firm level, whereas the separation by sectors clusters by sector (Fiechter et al., 2022). These considerations result in equations ((6) and (7)) and ((8) and (9)) to answer ***H2***:

For LowESG firms:

$$\begin{aligned} ESG_{it}^{Low} = {} & \beta_0^* + \beta_1^*(Treatment_i * Mandate_t) \\ & + \beta_2^* \ln(SIZE)_{it} + \beta_3^* LEV_{it} \\ & + \alpha_i + \mu_t + \gamma_{jt} + \varepsilon_{it} \end{aligned} \tag{6}$$

$$\begin{aligned} Q_{it}^{Low} = {} & \vartheta_0^* + \vartheta_1^* \widehat{ESG}_{it}^{Pre} + \vartheta_2^* \ln(SIZE)_{it} \\ & + \vartheta_3^* LEV_{it} + \alpha_i + \mu_t + \gamma_{jt} + \varepsilon_{it} \end{aligned} \tag{7}$$

For HighESG firms:

$$\begin{aligned} ESG_{it}^{High} = {} & \beta_0^{**} + \beta_1^{**}(Treatment_i * Mandate_t) \\ & + \beta_2^{**} \ln(SIZE)_{it} + \beta_3^{**} LEV_{it} \\ & + \alpha_i + \mu_t + \gamma_{jt} + \varepsilon_{it} \end{aligned} \tag{8}$$

$$\begin{aligned} Q_{it}^{High} = {} & \vartheta_0^{**} + \vartheta_1^{**} \widehat{ESG}_{it}^{Post} + \vartheta_2^{**} \ln(SIZE)_{it} \\ & + \vartheta_3^{**} LEV_{it} + \alpha_i + \mu_t + \gamma_{jt} + \varepsilon_{it} \end{aligned} \tag{9}$$

Since sector affiliation tends to be associated with lower or higher ESG scores, cross-sectional hypothesis ***H2a*** is checked to obtain more robust results on ***H2a.*** It can be expected that sector affiliation is, on average value relevant following the passage of the NFRD (Grewal et al., 2019; Ioannou & Serafeim, 2017; Jackson et al., 2020; Krueger, 2015a; Mittelbach-Hörmanseder et al., 2021). For example, Krueger (2015a) shows that some industries have a more substantial environmental impact than others, such as the Oil and Gas and the Basic Materials industries. These firms are under higher public pressure and tend to release more ESG information to legitimize their actions, as research suggests (Cahan et al., 2016; Cho, Michelon, et al., 2015; Patten, 1992). Accordingly, to answer ***H2b,*** the value relevance for TRBC sectors is tested by introducing a second partition ("*ESGSector*"). This partition was created after the treatment group was ranked by mean ESG scores per industry based on the pre-directive matching year data from 2013 (see Table 1). The top three ESG score-means belonged to the sectors: Real Estate, Consumer Non-Cyclicals, and Healthcare ("HighESGSector"). The bottom three performing sectors were Basic Materials, Energy, and Industrials ("LowESGSector") ((Krueger, 2015a) reports similar bottom sectors). Firms in these sectors are used with their previously matched U.S. control firm. Followingly, similar to equations (6) to (9), the IV2SLS are run for both sub-groups individually.

6. Empirical Results

6.1. Descriptive Statistics

Table 2 Panel A and B report the descriptive statistics for the dependent and independent variables of the treatment and the control group after matching. With the mean (M) of the ESG score and each pillar score lower than the median (50%), a slightly left-skewed distribution can be observed for the scores in the treatment and the control group. Opposing the left-skewness, Panel A and B report a slight right-skewness throughout both groups for the control variables ln(SIZE), Tobin's Q (Q), and leverage (LEV) as well as for the firm characteristics revenue, assets, and ROA (variables defined in *Appendix A*). Moreover, the mean of the individual social score is higher than that of the environmental and governance score, which are relatively similar to the treatment group. The control group has a higher mean governance score than the social and even higher than the environmental score.

The average firm in the treatment group has an ESG score of 54.366 and relatively similar Environmental and Governance scores of 51.316 and 51.351, respectively, while the Social score is higher with a mean of 57.624. The firm reports a natural logarithm of total assets of 32.145, Tobin's Q of 1.92, leverage of 0.6, revenues of around EUR€ 10.05mn., total assets of EUR€ 14.74mn., and a ROA of 5.903% on average. In comparison, the control group can put forward a mean for the ESG score of 54.769. Compared to the treatment group, the Social score is marginally lower, whereas with a mean of 48.2, the Environmental score has a more pronounced distance to its EU peer. Opposite, the Governance score is significantly higher for the average firm in the control group, with a mean of 58.3. The average control firm reports a natural logarithm of total assets of 32.707 with a Tobin's Q of 2.338 and leverage of 0.631. Firm characteristics not used for matching are higher for the average control firm with revenues of EUR€ 15.15mn., total assets of EUR€ 18.28mn., and a ROA of 7.114%. A country and sector distribution for the sample after matching is included in Chapter 4.4 *Summary Statistics* in Table 1 Panel B, C, and D.

Table 2: Descriptive Statistics

Panel A: Descriptive Statistics of Dependent and Independent Variables and General Characteristics After Matching - Treatment Group

	ESG Score	Env. Score	Soc. Score	Gov. Score	ln(SIZE)	Q	LEV	Revenue	Assets	ROA
Firm Years	3139	3139	3139	3139	3139	3139	3139	3139	3139	3139
M	54.316	51.836	57.624	51.351	32.145	1.920	0.600	10,052,070,000	14,735,840,000	5.903%
Min	0.627	0.000	0.432	0.867	26.440	0.250	0.026	0	91,053,470	-33.262%
25%	41.386	32.563	39.969	33.504	30.590	1.114	0.480	1,219,148,000	1,616,346,000	2.432%
50%	54.784	54.117	59.081	52.626	31.994	1.478	0.595	3,135,515,000	4,276,185,000	5.050%
75%	68.638	73.481	76.331	68.842	33.492	2.176	0.719	9,333,205,000	12,078,360,000	8.818%
Max	94.688	97.435	98.185	95.993	38.737	18.490	1.812	292,277,000,000	458,156,000,000	100.620%
SD	18.668	25.810	22.866	21.669	2.100	1.454	0.197	22,393,940,000	33,709,640,000	7.532%

Panel B: Descriptive Statistics of Dependent and Independent Variables and General Characteristics After Matching - Control Group

	ESG Score	Env. Score	Soc. Score	Gov. Score	ln(SIZE)	Q	LEV	Revenue	Assets	ROA
Firm Years	3157	3157	3157	3157	3157	3157	3157	3157	3157	3157
M	54.769	48.200	56.170	58.300	32.707	2.338	0.631	15,151,870,000	18,277,690,000	7.114%
Min	6.866	0.000	0.840	1.032	27.100	0.296	0.065	38,545,500	143,824,900	-54.915%
25%	40.682	25.975	39.171	43.367	31.314	1.371	0.497	2,252,896,000	2,669,145,000	3.423%
50%	55.696	52.228	57.225	60.223	32.546	1.876	0.612	4,984,658,000	6,269,581,000	6.611%
75%	69.522	71.054	73.935	74.452	33.949	2.763	0.749	12,876,730,000	16,587,390,000	10.571%
Max	95.162	98.546	98.011	98.795	38.755	15.969	2.919	450,088,500,000	463,742,900,000	58.026%
SD	18.624	26.984	21.540	20.480	1.890	1.520	0.222	34,599,310,000	37,938,420,000	7.348%

Chapter 4.3 reports an explanation for the differences in the amount of firm years in the treatment and control group. General firm characteristics include Revenue, Assets, and ROA. Revenue and Assets are reported in EUR€. "M" stands for "Mean", "Min" is the minimum value, "25%, 50%, and 75%" report quartiles, "Max" is the maximum value. "SD" refers to the standard deviation. All variables are defined in Appendix A.

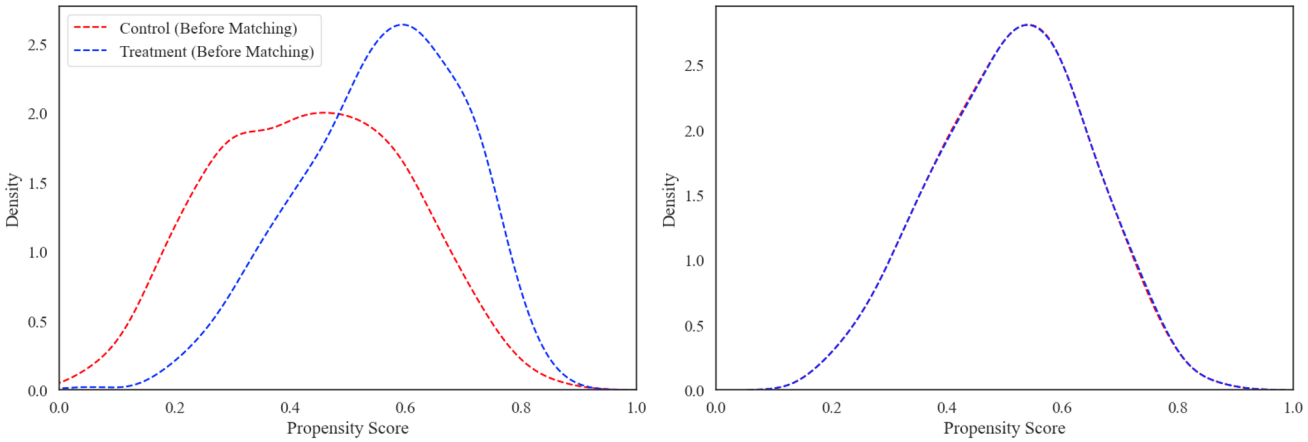


Figure 1: Distribution of Propensity Scores Before (Left) and After (Right) Matching

Table 3 and corresponding Figure 1 show the results of the propensity score matching on the used variables. These variables stay below the threshold of 0.25 standardized difference of means suggested by Rubin (2001), which indicates trustworthy regression adjustments. This is further supported by the concentration of the matched sample near the 0.5 propensity score (Figure 1) and no further distribution into extremes in the 0 or 1 direction, which would indicate a very high or low probability of falling under the treatment condition (F. Li & Thomas, 2018).

Table 3: Standardized Mean Differences Across Covariates Before and After Matching

Matching Variable	Effect Size in Standardized Mean Differences	
	Before Matching	After Matching
ln(SIZE)	0.139	0.240
LEV	0.132	0.063
ESG Score	0.497	0.128
Q	0.126	0.182
Sector_Basic Materials	0.155	0.008
Sector_Consumer Cyclicals	0.004	0.019
Sector_Consumer Non-Cyclicals	0.027	0.044
Sector_Energy	0.008	0.037
Sector_Healthcare	0.130	0.072
Sector_Industrials	0.185	0.035
Sector_Real Estate	0.165	0.050
Sector_Technology	0.140	0.025

The shown table depicts the standardized mean differences across all matching variables. All data is retrieved via Refinitiv Eikon Data API. PSM is conducted based on the pre-directive values of the matching variables in 2013 and uses a caliper of 0.05 without replacement. Variables used are: ln(SIZE), Sector association, Leverage, ESG score, and Tobin's Q.

6.2. Baseline Analysis

Aggregated ESG (H1a) and Individual E, S, and G (H1b) Firm Value Effects

To answer the research question drawn out at the beginning of this work on whether and if so, with which significance the introduction of the NFRD affected firm value, a two-stage IV approach was chosen in line with relevant scholarship (Allman & Won, 2022; Cahan et al., 2016; Fiechter et al., 2022; Gibbons, 2020; Ioannou & Serafeim, 2017). Following the notion that firm value is associated with ESG performance (e.g., Fatemi et al., 2018) and Refinitiv's ESG score may serve as a proxy for firm performance (e.g., Eccles et al., 2014; Fiechter et al., 2022), the first stage of the approach tests the legal shock of the NFRD for validity by exhibiting that ESG scores of treatment firms significantly increase after the passage of the NFRD compared to control firms.

Figure 2 depicts the development of the average ESG Scores for the matched sample. It shows the development from 2011 to 2019, with 2014 marked as shock date. The dotted lines show the continuation of the 2011-2014 trend. The figure demonstrates parallel trends leading up to the shock year 2014. From 2014 on, the average ESG score of the treatment group (EU firms) increases more significantly than the score of the control group (U.S. firms), which grows as well but starts later and with a significantly lower slope. A similar trend development for EU+ and U.S. firms between 2011 and 2018 can be found in Fiechter et al. (2022, p. 1516).

Uncovering the driving pillar of the ESG score development, *Appendix B* reports a comparison of the average Environmental, Social, and Governance scores, respectively, for the control and treatment group. Strikingly, the increase in the overall ESG score for the control group in 2015 was mainly driven by the Governance score, whereas for the treatment group, all three scores showed an increase from 2014 onwards. Since market participants are mainly interested in the combined ESG score – as explained in section 4.5 *Main Variables* – the utilization of the combined ESG score is continued. However, following this observation, a robustness test in the spirit of Ioannou and Serafeim (2017), which employs the Governance score as a further control variable for the primary 2SLS model, is included in the robustness test section.

The increase in ESG scores in the EU starts as early as 2014 when the NFRD was adopted, which supports the findings of Cuomo et al. (2022), Fiechter et al. (2022), Grewal et al. (2019), and Mittelbach-Hörmanseder et al. (2021) who uncovered the existence of anticipatory effects prior to the entry into force of the NFRD in 2017. As Table 2 and Figure 2 depict, treatment and control group show relatively similar ESG scores throughout the observation period, which reflects the employed matching variable ESG_i.

The mere graphical presentation of the sample ESG scores does not provide enough evidence to confirm the shock, as other confounding effects could lead to the increase. Thus, to increase the exploitability of the findings, the first stage of the IV2SLS works with controls for leverage and size, standard errors clustered at the firm level, and firm, year, and year-sector fixed effects. Table 4 provides results for the effect of the NFRD on the four dependent variables, ESG performance and the disaggregate E, S, and G scores, respectively. Results on the controls, the coefficient β_1 on the interaction term $Treatment_i * Mandate_t$, their respective significance and standard errors, and the fit of the estimated model are included. In the spirit of Fiechter et al. (2022), Ioannou and Serafeim (2017), and Ottenstein et al. (2022), models are reported with all three fixed effects included.

The results of running equation (2) on an aggregate and disaggregate level for the ESG and the E, S, and G score individually reveal a strong positive as well as highly statically significant effect of the NFRD on ESG performance as the coefficient on the interaction term $Treatment_i * Mandate_t$ displays with a value of 3.007 units on the aggregate level (Table 4). On a disaggregated level of the pillars, this effect is most pronounced and statistically significant at the 1%-Level of the Social score. This suggests that this pillar had the most substantial contribution to the overall ESG performance increase. Interestingly, this coincides with the findings of Ioannou and Serafeim (2017, p. 40), who found that "the magnitude of the effect on social disclosures seems to be larger than the respective effect on environmental or governance disclosures.". This finding has to be handled carefully, as Ioannou and Serafeim analyzed ESG disclosure mandates outside the European Union. Another relevant obser-

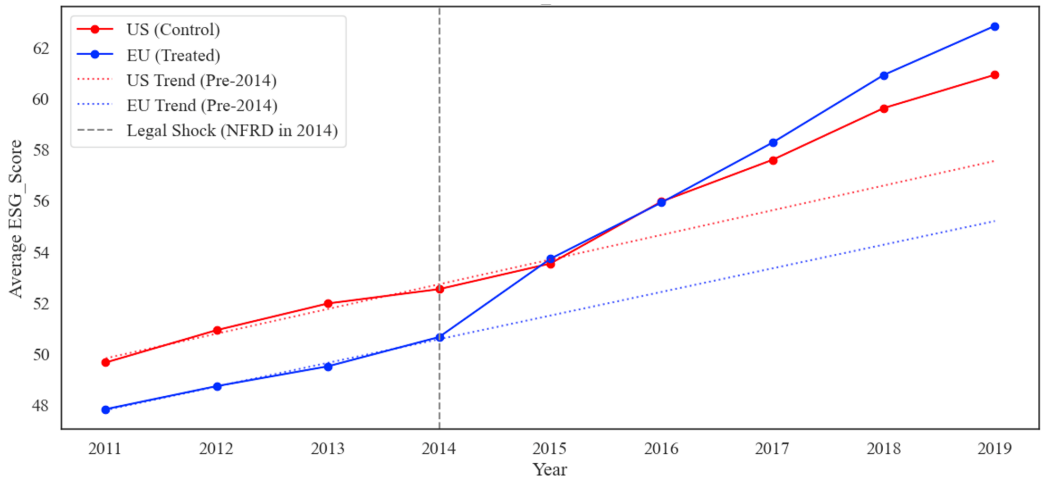


Figure 2: Comparison of Average ESG Scores Over Time with Pre-Directive Trend Lines

Table 4: First Stage: Impact of the NFRD on ESG Performance - (Dis)Aggregate Effects

	ESG Score	Env. Score	Soc. Score	Gov. Score
Intercept	6.468	-3.203	24.806	-8.298
	(16.887)	(24.744)	(21.540)	(25.749)
Treatment * Mandate	3.007***	0.392	5.912***	1.608
	(0.708)	(0.941)	(0.879)	(1.172)
ln(SIZE)	1.687***	1.827**	1.41**	1.98**
	(0.518)	(0.757)	(0.656)	(0.787)
LEV	0.012	-1.772	-1.841	3.155
	(1.851)	(2.379)	(2.144)	(3.252)
Firm fixed effects	Included	Included	Included	Included
Year fixed effects	Included	Included	Included	Included
Year-Sector fixed effects	Included	Included	Included	Included
Adjusted R^2	0.859	0.879	0.84	0.654
# of observations	6296	6296	6296	6296

The shown table depicts results of the coefficient on the interaction term of two dummy variables, $Treatment$ (1 if in treatment group (EU), 0 if in control group (U.S.)) and $Mandate$ (1 if after shock date in 2014 (2014-2019), 0 if before shock date (2011-2013)). This first stage of an IV2SLS is estimated in a difference-in-difference OLS regression using equation (2). Dependent variable of interest is ESG performance measured by ESG score and the three disaggregate pillar scores respectively. All retrieved via Refinitiv Eikon Data API. Control variables are: ln(SIZE) and LEV. ln(SIZE) is the natural logarithm of total assets. LEV is the total liabilities over total assets. ***, **, and * indicate statistical significance either at the 1%, 5%, or the 10% level (two-tailed), respectively. Following Flammer (2015) and Gibbons (2020), standard errors are included in parentheses. Standard errors clustered at the firm level are employed. The model was robust to running combinations of one less fixed effect of each. All variables are defined in Appendix A.

vation can be made for the positive coefficient on the variable size (ln(SIZE)), which is significant at the 1%-level across all specifications. As prior research found, larger firms are inclined to release more ESG-related information and have a higher ESG performance and scores (Baran & King, 2014; Drempetic et al., 2020; M. Lang & Lundholm, 1993). The results that equation (2) and its adapted version for the three individual scores yield support the findings of prior research that larger firms tend to have a higher ESG performance and score.

After confirming the shock to ESG performance in the first stage and estimating results on the interaction term, the second stage of the 2SLS model utilizes the predicted $\widehat{ESG}_{it}$ from the first stage to explore the effect of the NFRD-induced change in ESG performance on the dependent variable Tobin's Q as a measure of firm value. The model covers approximately 80% of the variance of Tobin's Q. In ***Hypothesis 1a,*** a negative capital market reaction via ESG scores, condensing into a low negative effect on firm value following the introduction of the NFRD, was formulated. This hypothesis can be

confirmed following an IV2SLS analysis on a propensity score matched sample of 354 U.S. and 354 EU firms. Indeed, statistically significant findings on the coefficient of "instrumented score" tabulated in Table 5 suggest that firm value (Tobin's Q) decreases after the introduction of the NFRD. Using the predicted $\widehat{ESG}_{it}$ from the first stage (see equation (2)), a statistically significant (at the 1%-level) negative value reaction of -0.078 on the instrumented ESG score can be reported for the aggregate effect (Table 5). Thus, firms in the treatment group experienced an average 0,078 unit decrease in firm value after the introduction of the NFRD from 2014-2019 compared to the control group of U.S. firms.

The disaggregate instrumented scores on E, S, and G show that each of the three pillars influenced Tobin's Q negatively but with a different magnitude, confirming ***Hypothesis 1b***. Interestingly, the social score, the only statistically significant and most substantial reactor in stage one, is the lowest of the three scores in stage two. Out of the three individual scores, the Environmental score reacts the strongest with a negative value of 0.598 (statistically significant at the 1%-level) after the introduction of the NFRD. In line with the results brought forward by Ioannou and Serafeim (2017), the coefficient on the control variable ln(SIZE) shows a negative sign throughout all four scores except for the Environmental score, which is positively associated. Results on the coefficient on LEV suggest a positively associated nature throughout all four scores, again, except for the Environmental score.

An explanation for this could potentially be the lower prevalence of topics that had a more pronounced negative market reaction in the annual or standalone reports. With the mandated disclosure of "information [...] relating to, as a minimum, environmental, social and employee matters, respect for human rights, anti-corruption and bribery matters..." (Directive 2014/95/EU (2014a), Article 19a (1)), and thus those topics that are included in e.g. the Environmental score which drove most of the negative market reaction, less prevalent topics are disclosed through the directive and thence lead to a more significant market reaction (Mittelbach-Hörmanseder et al., 2021).

The tabulated results on the disaggregate effects show – except for the contrasting sign for all values – a pattern tying with results previously reported by Ioannou and Serafeim (2017) on a worldwide and a U.S. control group. The Environmental score features the highest effects, followed by the Governance and Environmental scores. Moreover, the results from Table 5 match those of Mittelbach-Hörmanseder et al. (2021) first, in the overall negative market reaction to the NFRD for ***H1a*** and secondly, for ***H1b***, in the more substantial impact of the environmental score compared to the social score (Mittelbach-Hörmanseder et al., 2021, p. 325). Additionally, verifying the product of research of Fiechter et al. (2022), who tabulated a negative albeit not statistically significant firm value reaction to the adoption of the NFRD of -0.048, the IV analysis in the work at hand obtains results with a similar magnitude of -0.078 (Table 5).

As laid out in Chapter 6.3 on prior research in the field of mandatory ESG disclosure, scholarship on firm value implications of ESG disclosure mandates in countries outside the EU brings forward mixed but mainly positive market reactions (negative: Chen et al., 2018; positive: Ioannou and Serafeim, 2017; Krueger, 2015a; Lee and Yeo, 2016; T. Li et al., 2023). In comparison, the relatively small and emergent strand of scientific literature focussing on a specific disclosure mandate, the NFRD in Europe, the three studies mentioned in the literature review suggest a negative equity market perception of the mandated ESG disclosure (Fiechter et al., 2022; Grewal et al., 2019; Mittelbach-Hörmanseder et al., 2021).

The findings of ***Hypothesis 1a*** underpin existing literature on negative firm value implications of the NFRD. Additionally, in line with previous literature, the results of ***Hypothesis 1b*** show that, on average, the mandated ESG disclosure has a differentiated topical relevance condensation into a mixed market reaction depending on the topic disclosed, hence supporting ***H1b***.

Early and Late Post-Period Firm Value Effects (H1c)

Tabulated empirical results for ***Hypothesis 1a*** and***1b*** show that not only did treatment group firms' ESG performance increase after the adoption of the NFRD in 2014, but also firm value was negatively impacted. Based on these merely general, *post-directive* period results spanning from 2014 to 2019, one cannot identify whether the effects manifested with anticipatory effects or simply after the entry into force of the directive in 2017. Therefore, to assess ***Hypothesis 1c***, whether the effects of the NFRD on firm value are more significant in the late post-period (2018-2019) than in the early post-period (2014-2017), in line with the approach of Fiechter et al. (2022) yearly effects are estimated in a two-stage DID estimation, using equations (4) and (5). The predicted ESG score from the first stage does not function as in the IV but merely posits controlling properties in this setting in the second stage. Tables 6 and 7 report results on the first and second stage of the yearly effects estimation. Figures 3 and 4 graphically visualize the treatment point effects in a confidence band diagram, respectively.

The results in the first stage show that on average, in the years after the passage of the NFRD, firms' ESG performance significantly increased. Based on the findings in Table 6, which reveal statistically significant increasing coefficients on the yearly interaction terms for 2015 to 2019 with the base year 2013 (thence excluded), it becomes clear that after the European Union adopted the NFRD in 2014, anticipatory effects kicked in and ESG performance measured by ESG scores increased for the years mentioned. The most significant year out of the nine years under observation is 2015, with a value of 2.306, which is around 2.7 times bigger than the second highest value, 0.843 in 2017. This suggests that the anticipation of the NFRD by far had the single most significant impact on the ESG performance of firms rather than the actual entry-into-force of the directive.

Table 5: Second Stage: Impact on Tobin's Q - Aggregate and Disaggregate Effect

	ESG Score	Env. Score	Soc. Score	Gov. Score
Intercept	19.048***	16.532***	19.612***	16.948***
	(1.911)	(1.867)	(1.955)	(1.857)
Instrumented score (respect.)	-0.078***	-0.598***	-0.04***	-0.146***
	(0.022)	(0.166)	(0.011)	(0.040)
ln(SIZE)	-0.335***	0.627**	-0.411***	-0.178**
	(0.063)	(0.296)	(0.059)	(0.089)
LEV	0.857**	-0.205	0.783**	1.316***
	(0.375)	(0.445)	(0.373)	(0.412)
Firm fixed effects	Included	Included	Included	Included
Year fixed effects	Included	Included	Included	Included
Year-Sector fixed effects	Included	Included	Included	Included
# of observations	6296	6296	6296	6296

The shown table depicts results of the coefficient on the Predicted ESG score and the three disaggregate pillar scores from stage 1 respectively. Dependent variable of interest is Tobin's Q. This second stage of an IV2SLS, uses equation (3). All input retrieved via Refinitiv Eikon Data API. Control variables are: ln(SIZE) and LEV. ln(SIZE) is the natural logarithm of total assets. LEV is the total liabilities over total assets. Tobin's Q is calculated as sum of assets and market value of equity minus common book equity over the total assets. Adjusted R^2 is at 0.799. ***, **, and * indicate statistical significance either at the 1%, 5%, or the 10% level (two-tailed), respectively. Following Flammer (2015) and Gibbons (2020), standard errors are included in parentheses. Standard errors clustered at the firm level are employed. All variables are defined in Appendix A.

Grounded on the indications in Table 6, the anticipatory nature of the findings is in line with the findings of Fiechter et al. (2022) and Grewal et al. (2019), who were among the first to report that companies already in 2014 started to adapt to the upcoming entry-into-force of the directive in 2017. Fiechter et al. report an increase in ESG performance from 2014 on, but just in 2016, this increase becomes statistically significant. In congruence with their findings, compared to the U.S. control group, no meaningful pre-directive trends can be observed, as coefficients on the interaction term for 2011 and 2012 are low and insignificant.

Results are further subjected to graphical inspection (Figure 3), highlighting 2015 as the first year with statistical significance and the highest increase to the prior year. Similar to Fiechter et al., 2022, who report the first significant increase for the Refinitiv ESG score in 2016 and the subcomponent Social score in 2015. The jump in ESG scores in 2015 compared to 2014, the succeeding stable increase in 2016, and a following surge in 2017 again may be caused by general public attention in the years around the directive's adoption (Jessop & Kelly, 2014). Further, changes in management behaviour following stakeholder pressure or increased benchmarking opportunities to peers (Table 6) (Chen et al., 2018; Christensen et al., 2021; Fiechter et al., 2022; Ioannou & Serafeim, 2017; Tomar, 2019). Either, companies belong to the group that prepares early for the directive and plans on staying ahead of its competitors and have the "first-mover" advantage, which the capital market might reward (also see "race-to-the-top" by Ioannou and Serafeim (2017, p. 5 and 23). Alternatively, the company plans on waiting until the very last moment before mandatory compliance to increase efficiency in case any changes to the mandate or within the company happen until the entry into force ("internal learning" according to Fiechter et al., 2022, p. 1516). Moreover, Fiechter et al. (2022) suggest that even in the absence of immediate stakeholder pressure in the early years (2014-2015), an anticipation of stakeholder reactions could be a driver for the increase in ESG scores as it is harder to restore the lost reputation of clients and investors than to preserve it (Bolton et al., 2021; Fiechter et al., 2022). Further research might uncover the underlying drivers for these hypothesized explanations. From 2014 to 2017, an average 0.932 increase per year can be observed for the early post-directive period, whereas only 0.527 units of growth are tabulated for the late post-directive period. This described development of the ESG performance is plotted in Figure 3 with the accumulated yearly treatment effects from 2011 to 2019 for the dependent variable ESG score.

The results of the first stage in Table 6 are, on average, lower than what e.g., Ioannou and Serafeim (2017, p. 40) reported in their work, where they assessed mandatory disclosure regulations in four countries around the world and outside of the European Union in comparison to a worldwide and U.S. control group. At the same time, the results in Table 6 are, on average, higher than what Allman and Won (2022) and Fiechter et al. (2022) estimate in their works assessing the NFRD. Therefore, the magnitude of the findings in this work is within reasonable boundaries compared to results from prior research.

The second stage of the yearly effects model utilizes a DID equation with the predicted ESG score from the first stage as

Table 6: Impact of the NFRD on ESG Performance - Yearly Effects

Dependent Variable		Accumulated ESG Score effect	Yearly effects
*2011*Treatment*		0.222	0.153
		(0.696)	
*2012*Treatment*		0.069	0.069
		(0.506)	
*2014*Treatment*	Adoption	0.571	0.571
	of NFRD	(0.466)	
*2015*Treatment*		2.877***	2.306***
		(0.626)	
*2016*Treatment*		2.885***	0.008***
		(0.736)	
*2017*Treatment*		3.728***	0.843***
		(0.866)	
*2018*Treatment*	Entry-into-force	4.072***	0.344***
		(0.962)	
*2019*Treatment*		4.781***	0.709***
		(0.981)	
Controls	Included		
Firm fixed effects	Included		
Year fixed effects	Included		
Year-Sector fixed effects	Included		
Adjusted R^2	0.86		
# of observations	6296		

The shown table depicts yearly effects estimated in a difference-in-difference OLS regression using equation (4). Dependent variable of interest is ESG performance measured by ESG score retrieved via Refinitiv Eikon Data API. Year 2013 is used as base year and therefore excluded. Control variables are: ln(SIZE) (1.814***, 0.520)and LEV (0.638, 1.854). ln(SIZE)is the natural logarithm of total assets. LEV is the total liabilities over total assets. ***, **, and * indicate statistical significance either at the 1%, 5%, or the 10% level (two-tailed), respectively. Following Flammer (2015) and Gibbons (2020), standard errors are included in parentheses Standard errors clustered at the firm level are employed. The model was robust to running all combinations of fixed effects. All variables are defined in Appendix A.

a control variable to estimate the yearly treatment effect on Tobin's Q as the measure of firm value. Table 7 reports findings on the accumulated and yearly effects of the NFRD on the interaction terms of Year*Treatment for years 2011-2012 and 2014-2019 since 2013 is denoted as the base year. These findings are presented graphically in Figure 4. The model is run with the exact specifications as equations (2) and (3) for the IV model and equation (4) for the first stage, except for the newly included predicted ESG score from the first stage (equation (4)) Coefficients on the controls are reported in the note under Table 7.

As the OLS regression utilized in this setting simultaneously serves as an early robustness test to the findings answering ***H1a*** and ***H1b***, results in Table 7 suggest that, on average, the negative market reaction to mandatory ESG disclosure in the EU can be confirmed, which further strengthens the implications of this work. Years after the adoption showed, in comparison to the two years prior to the base year 2013, a significant decrease in firm value on average. Out of the six years, two (2015 and 2017) demonstrate positive signs, whereas market responses are negative with different magnitudes in all other years after 2013. Similar to the results (Table 6) in stage one, the first significant coefficient on the interaction term after the introduction of the NFRD is the most considerable absolute value. Instead of 2015 for the first stage with ESG performance as the dependent variable, it is 2014 with Tobin's Q as the dependent variable, which reports the most considerable negative value (-0.287), followed by 2016 with a negative reaction of -0.227. Both values are statistically significant at the 1%-level.

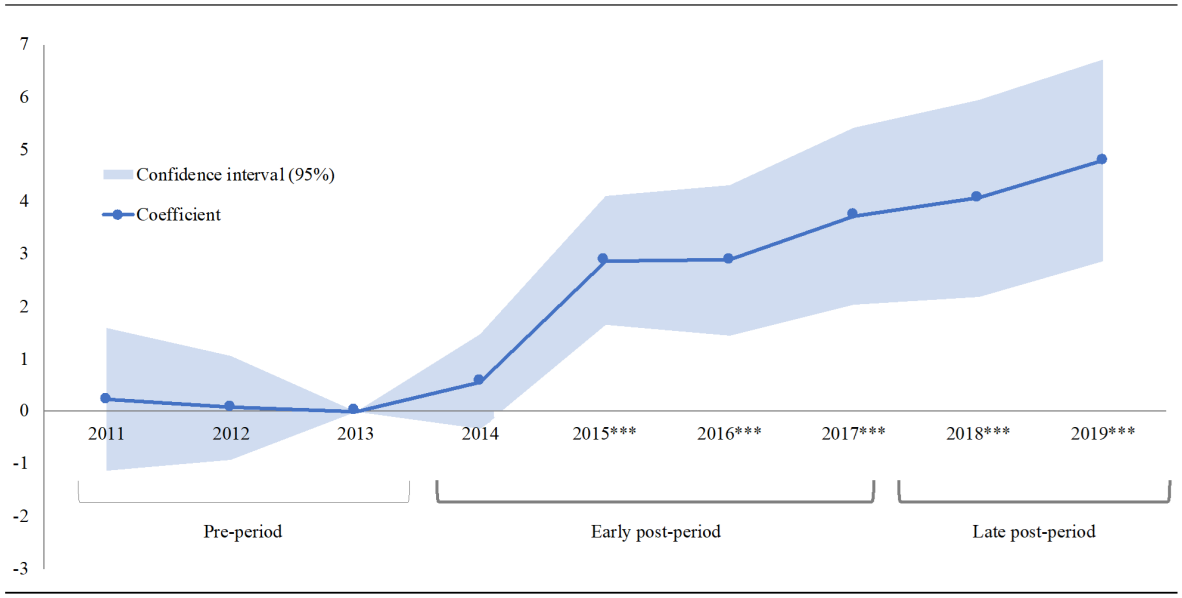


Figure 3: Yearly Treatment Effects – ESG score – with 95% Confidence Interval

The shown figure depicts the yearly treatment point estimates of ESG performance in a band plot. The dark blue line shows the yearly effects. The lighter blue shows the upper and lower boundaries of the 95% confidence interval. Year 2013 is used as base year and therefore excluded (shown as 0). This first stage of the yearly effects is estimated in a difference-in-difference OLS regression using equation (4). Dependent variable of interest is ESG score (as a proxy for ESG performance) retrieved via Refinitiv Eikon Data API. ***, **, and * indicate statistical significance of the interaction term either at 1%, 5%, or the 10% level (two-tailed), respectively for each year. All variables are defined in Appendix A.

Potentially, the negative market reaction in 2016 came from the Paris Agreement adopted by 196 Parties at the UN Climate Change Conference in Paris on December 12, 2015, and entered into force on November 4, 2016. This landmark event heaved worldwide commitment to ESG and thus reinforced the importance and impact of the NFRD (Hummel & Jobst, 2022; UNFCCC, n.d.). This relevance of the year 2016 falls into line with the novel contributions of Gibbons (2020, updated April 2023) who analysed detailed holdings data for firms subject to the staggered introduction of 40 country-level regulations which mandate disclosure and found that, generally, mandatory non-financial disclosure did not reduce ESG-based adverse selection problems in his sample and thus had no value relevance. Instead, he reported that the institutional clientele channel influenced value. Following mandatory non-financial disclosure, Gibbons' findings suggest that material real effects for firms emerge because ESG disclosure attracts institutional investors with ESG preferences and longer-term investment horizons. A landmark event such as the Paris Climate Conference could potentially bring elevated attention and relevance to the topic of ESG and thus to the NFRD from an institutional investor's perspective, which resulted in an increased focus on ESG in investment decisions in the EU that lead to the negative reaction in 2016 among two years of low positive value effects (2015 and 2017). This research-backed public attention channel was already mentioned in a previous explanation for the increase in ESG scores (Christensen et al., 2017, 2021; Fiechter et al., 2022; Grewal et al., 2019). The potential existence of this value-destruction channel may be attributable to the clientele-channel mechanism but could also be due to a reduction in adverse selection, following, e.g., the findings of Krueger (2015a), who showed that in a GHG emission disclosure mandate in the UK, this adverse selection problem was reduced. Indeed, these findings do not explain a positive or negative value relevance of the NFRD but explain why 2016 stands out (Gibson & Krueger, 2017; Liang & Renneboog, 2020).

Adducing the Efficient Market Hypothesis and the knowledge of the sample composition, which includes, on average, only large stocks on established markets, the findings from Table 7 indicate that as soon as the adoption of the NFRD happened, market participants priced in the anticipated costs (benefits) of the NFRD, which furthermore explains the flip from positive to negative sign from 2017 to 2018, the years of entry-into-force (Fama, 1970). This was the first year when prior non-reporting firms subject to the thresholds were forced to disclose ESG information. Coming back to the results of Table 6, the first significant and positive ESG performance effects manifested in 2015, which shows the difference in reaction speed: Markets reacted immediately (effect in 2014), whereas firms needed to adapt their ESG actions and ESG scores consequently needed adjustment (effect in 2015). A similar conclusion was reached by Grewal et al. (2019), who analysed the market reaction to events leading up to the NFRD and found a negative market reaction to the adoption of the NFRD as well.

Even though the market reacted as early as 2014 to the adoption of the NFRD, the for the subclusters' early and late post-directive periods, on average, most significant negative

Table 7: Impact of the NFRD on Tobin's Q - Yearly Effects

Dependent Variable		Accumulated Tobin's Q effect	Yearly effects
*2011*Treatment*		-0.133**	-0.163**
		(0.059)	
*2012*Treatment*		0.032	0.032
		(0.048)	
*2014*Treatment*	Adoption	-0.287***	-0.287***
	of NFRD	(0.053)	
*2015*Treatment*		-0.162**	0.125***
		(0.077)	
*2016*Treatment*		-0.389***	-0.227***
		(0.076)	
*2017*Treatment*		-0.334***	0.055***
		(0.084)	
*2018*Treatment*	Entry-into-force	-0.489***	-0.155***
		(0.096)	
*2019*Treatment*		-0.669***	-0.18***
		(0.119)	
Controls	Included		
Firm fixed effects	Included		
Year fixed effects	Included		
Year-Sector fixed effects	Included		
Adjusted R^2	0.8		
# of observations	6296		

The shown table depicts yearly effects estimated in a difference-in-difference OLS regression using equation (5). Dependent variable of interest is Tobin's Q calculated with items retrieved via Refinitiv Eikon Data API. Year 2013 is used as base year and therefore excluded. Control variables are: ln(SIZE) (-0.539***, 0.067), Predicted ESG Score (0.039***, 0.004), and LEV (0.801**, 0.375). ln(SIZE)is the natural logarithm of total assets. LEV is the total liabilities over total assets. Tobin's Q is calculated as sum of assets and market value of equity minus common book equity over the total assets. Predicted ESG Score is predicted in the first stage regression of the yearly effects model. ***, **, and * indicate statistical significance either at the 1%, 5%, or the 10% level (two-tailed), respectively. Following Flammer (2015) and Gibbons (2020), standard errors are included in parentheses. Standard errors clustered at the firm level are employed. The model was robust to running all combinations of fixed effects. All variables are defined in Appendix A.

reaction happened in the late post period from 2018 to 2019 after the entry into force. This can be concluded from averaging the yearly effects within the early post-period (-0.084) and the late post-period (-0.168), respectively. These findings support the notion hypothesized in ***H1c*** that the negative firm value implications found in ***H1a*** are likely to be more significant in the late post-period (2018-2019) than in the early post-period (2014-2017). Nevertheless, this confirmation must be handled cautiously since the post-period merely entails results for two years, whereas the early post-period accounts for four years. Comparing two averages with a low number of years in each does not yield robust mean values.

As mentioned earlier, this work does not intend to shed light on the specific value creation or destruction channels with which the NFRD affects firm value. Questions like how market participants value the newly released information, how much the comply-or-explain-approach of the NFRD contributed to the value effects, or how the flaws in the construction of the directive are seen by markets are not answered here. As literature regarding the value implications of the NFRD is still in its infancy, solely an overall evaluation of the value effects of the directive is contributed with this work. Nevertheless, some minor suggestions for these relevant questions based on prior research are given to further back the findings of this work and explain their relevance. Further scholarship may pick up on the discussion of the underlying mechanisms and expand this field of research.

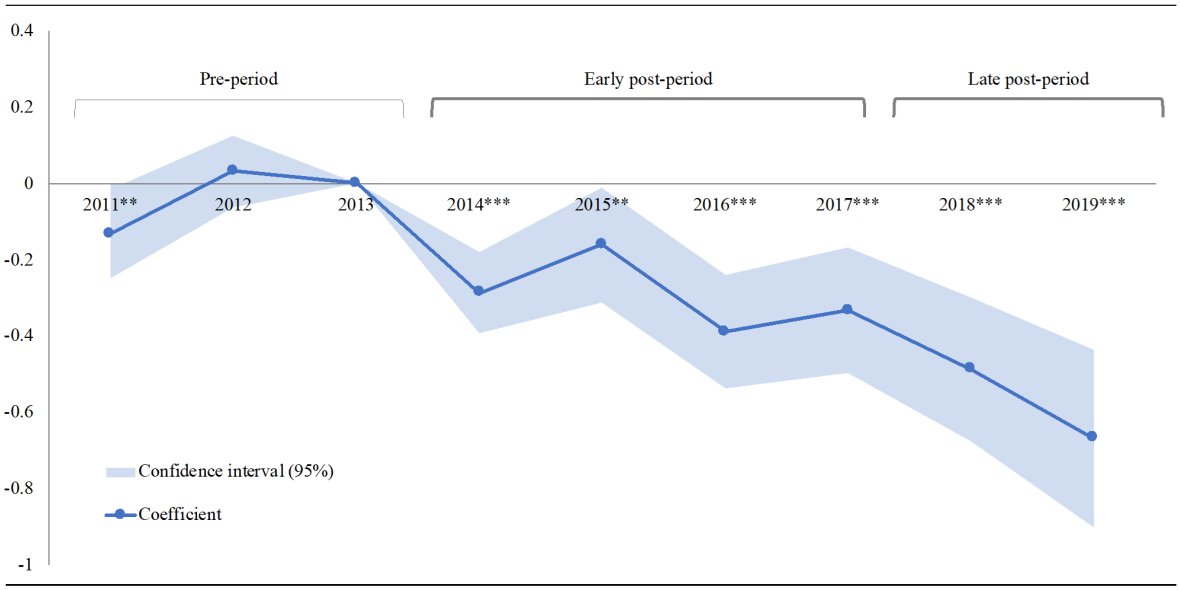


Figure 4: Yearly Treatment Effects – Tobin's Q – with 95% Confidence Interval

The shown figure depicts the yearly treatment point estimates of ESG performance in a band plot. The dark blue line shows the yearly effects. The lighter blue shows the upper and lower boundaries of the 95% confidence interval. Year 2013 is used as base year and therefore excluded (shown as 0). This first stage of the yearly effects is estimated in a difference-in-difference OLS regression using equation (5). Dependent variable of interest is ESG score (as a proxy for ESG performance) retrieved via Refinitiv Eikon Data API. ***, **, and * indicate statistical significance of the interaction term either at 1%, 5%, or the 10% level (two-tailed), respectively for each year. All variables are defined in Appendix A.

6.3. Cross-Sectional Analysis

Next, credible cross-sectional variation is investigated to answer ***H2a*** and ***H2b***. Grounded on prior research and the extant theoretical background given, it is to be expected that market reaction to the NFRD differs depending on prior ESG performance and sector association.

Results of running the two introduced cross-sectional partitions separately based on first, pre-directive high or low ESG disclosure and second, sector association with a partition into the three highest and three lowest ESG score sectors are reported in Tables 9 and 10. Both tables report combined results for the sectoral and pre-directive ESG performance relevance. Table 9 exhibits results for equations (6) and (7) with both stages of the IV2SLS done for each partition on *LowESG* and *LowESGSector*. Table 10 displays findings for equations (8) and (9) with, again, both stages of the IV2SLS but now for the higher performing partitions *HighESG* and *HighESGSector*. The sector partition (Table 8) includes the top three ESG score-mean sectors, Real Estate, Consumer Non-Cyclicals, and Healthcare ("HighESGSector") and the bottom three performing sectors, Basic Materials, Energy, and Industrials ("LowESGSector"). Low and HighESG was partitioned based on the top and bottom tercile of ESG firm performance in the matching year. The only conceptual alteration between the two models for sectoral and pre-directive ESG performance differences was made in the clustering of the standard errors, whereas for the sectoral model, the clustering was switched from firm to sector level.

As there is considerable overlap (conceptually and empirically) between the two partitions of *ESG* and *ESGSector*, not surprisingly, Table 9 exhibits similar results for LowESG and LowESGSector (first and second stage) and Table 10 for HighESG and HighESGSector (first and second stage). This was to be expected following extant literature confirming that some sectors have an on average higher or lower ESG score (Grewal et al., 2019; Ioannou & Serafeim, 2017; Jackson et al., 2020; Krueger, 2015a; Mittelbach-Hörmanseder et al., 2021).

Results in Table 9 expose that for the sample employed, LowESG and LowESGSector firms' coefficient (β_1^*) on the interaction term significantly (at the 1%-level) increases. Hence, firms in these groups post an increase in ESG performance after the introduction of the NFRD in 2014, with 3.996 and 4.127 units, respectively. For the HighESG and HighESGSector firms in Table 10, this increase in ESG performance does not turn out as severe as for the lower group in Table 9 but is still positive (2.792 and 2.622, respectively). In accordance with Ioannou and Serafeim (2017) and Jackson et al. (2020), who reported identical outcomes, the results allow for the educated deduction that the higher increase by lower-performing firms is caused by their overall lower level of ESG performance. Grounded on the theory laid out in the theoretical foundation part of this work, firms will only increase their disclosure if the marginal cost of every additional unit of disclosure is exceeded by the resulting benefits. Thence, firms with a lower level of ESG performance can more efficiently increase their performance by larger margins than companies with high ESG performance. Still, high ESG firms increase their disclosure as well, just not by the same magnitude. The reaction might be triggered by the

Table 8: Sector Partition Statistics - HighESGSector and LowESGSector

Sector (EU)	N	Firm Years	Firm years (%)	ESG Score	Env. Score	Soc. Score	Gov. Score	ln(SIZE)	Q	Lev
HighESGSector	82	731	23.288%	58.419	59.859	61.901	50.720	32.501	1.967	0.566
Real Estate										
Consumer Non-Cyclicals										
Healthcare										
LowESGSector	140	1245	39.662%	52.428	51.053	54.717	50.997	32.272	1.443	0.589
Basic Materials										
Energy										
Industrials										

The shown table depicts the partition in HighESGSector and LowESGSector. The split is based on the top three and lowest three sectors in terms of sectoral ESG performance in the sample in 2013. Measures "N" and "Firm Years" are summarized. All other measures reports means of the three indented sectors belonging to the partition.

Table 9: Impact of the NFRD on ESG Performance and Tobin's Q - LowESG(Sector)

	LowESG		LowESGSector	
	First Stage (ESG)	Second Stage (Q)	First Stage (ESG)	Second Stage (Q)
Intercept	32.84	19.039***	-1.407	10.26***
	(23.769)	(3.367)	(18.672)	(2.405)
coefficient 1 on β or ϑ	3.996***	-0.038	4.127***	-0.028***
	(1.279)	(0.030)	(0.325)	(0.009)
ln(SIZE)	0.316	-0.456***	1.784***	-0.23***
	(0.758)	(0.111)	(0.607)	(0.069)
LEV	1.995	1.044**	-2.921	0.495
	(3.189)	(0.514)	(2.755)	(0.339)
Firm fixed effects	Included	Included	Included	Included
Year fixed effects	Included	Included	Included	Included
Year-Sector fixed effects	Included	Included	Included	Included
Adjusted R^2	0.791	0.79	0.856	0.817
# of observations	2076	2076	2497	2497

The shown table depicts results of the two stages of the IV2SLS analysis conducted for both partitions to test H2a and H2b, LowESG and LowSectorESG. LowESG includes firms and their matched controls who have been in the bottom tercile of ESG performance in the year of matching 2013. LowSectorESG includes firms who are associated with the bottom three ESG performance sectors, namely Basic Materials, Energy, and Industrials ("LowESGSector"). Coefficient 1 on β and ϑ is the main variable of interest. It includes either: the coefficent from equation (6) for the first stage of the IV, β which is the coefficient on the interaction term of two dummy variables, *Treatment* (1 if in treatment group (EU), 0 if in control group (U.S.)) and *Mandate* (1 if after shock date in 2014 (2014-2019), 0 if before shock date (2011-2013)). Or, it includes coefficient ϑ which is the coefficient on the predicted ESG score in the second stage. Dependent variable of interest is ESG performance measured by ESG score in the first stage and Tobin's Q in the second stage. All data retrieved via Refinitiv Eikon Data API. Control variables are: ln(SIZE) and LEV. ln(SIZE) is the natural logarithm of total assets. LEV is the total liabilities over total assets. ***, **, and * indicate statistical significance either at the 1%, 5%, or the 10% level (two-tailed), respectively. Following Flammer (2015) and Gibbons (2020), standard errors are included in parentheses. Standard errors clustered at the firm level are employed for the Low/HighESG partition. Standard errors clustered at the sector level are used for the Low/HighESGSector partition. The model was robust to running all combinations of fixed effects. All variables are defined in Appendix A.

"race-to-the-top mechanism, mentioned in the *Early and Late Post-Period Firm Value Effects (H1c)* analysis. Firms with high ESG performance prior to the NFRD want to distance themselves from competitors to signal their superior ESG performance to the market. With an increasing focus on ESG in investment decisions, this behaviour would be rational if the associated line of action yielded a positive NPV for the firm.

Turning to the conditional effects revealed by the 2SLS regressions of Tobin's q on ESG performance and the control variables in Tables 9 and 10, it is to be concluded that all four firm value reactions for high and low partitions in ESG performance and sector affiliation occur in relatively moderate negative digits. Between high and low categorisation, Tobin's Q is affected similarly. A more negative and statistically significant reaction is suggested by coefficient ϑ_1^{**} for the HighESG and HighESGSector partitions, respectively (-0.092 and -0.099), compared to insignificant -0.038 for LowESG and -0.028 – significant at the 1%-level – for LowESGSector.

The introduced partitions allow for a more nuanced examination of the findings of the baseline analysis. The confirmation of an overall weakly negative firm value reaction to the introduction of the NFRD in 2014 can be backed by the findings in Tables 9 and 10. Moreover, following the remarks in this section and the results stated in Tables 9 and 10, hypothesis 2b, that for some sectors the effect of the NFRD is more significant than for others, can be confirmed.

However, following the results described in the previous paragraph, hypothesis 2a must be rejected, as for firms with higher pre-directive ESG performance, the effect of the NFRD on firm value is *not* less pronounced than for firms with lower pre-directive ESG performance. It even more comes to light that based on the findings in Tables 9 and 10, firms within the HighESG or HighESGSector partition are more negatively affected by the NFRD. This finding adds to disagreeing literature that, on the one hand, suggests that lower pre-directive ESG performance results in higher (negative) firm value effects (e.g., Grewal et al. (2019) and Ioannou and Serafeim (2017)) and, on the other hand states that the negative association between ESG performance and firm value is more pronounced for firms with higher ex-ante ESG performance (e.g., Cho, Michelon, et al. (2015), Fatemi et al. (2018), and Mittelbach-Hörmanseder et al. (2021)).

Fatemi et al. (2018) assume that this more intense negative market reaction could result from firms utilizing ESG disclosure and, hence ESG scores as a vehicle to signal good ESG performance irrespective of their actual ESG performance. This argument gains strength with the "comply-or-explain-approach" utilized in the NFRD – and criticised by scholars and practitioners alike Christensen et al. (2021) – that allows companies to opt out of reporting specific topics if they provide a reasonable explanation for not reporting and weakens the directive (Cho, Laine, et al., 2015). Additionally, a possible explanation might be grounded in overinvestment signalling to markets. Engaging in ESG activities is likely to be associated with costs. Resultingly, higher ESG scores would signal higher costs and potentially overinvesting in ESG (Kim & Lyon, 2015). Furthermore, for these lower ESG-performing firms, the market reaction might come with a lower absolute value because investors expect that these firms will increase their ESG scores (as also shown in this research) and can thus increase the interest of other market participants, which would raise their value. Therefore, they do not lower their value expectation to the extent they do for HighESG and HighESGSector firms. An equally plausible consideration from an investor's view might be the reduction of costs needed to maintain a weak ESG performance by paying, e.g., penalties. (Chang et al., 2022; Grewal et al., 2019; Healy & Palepu, 2001; Jensen & Meckling, 1978; Watts & Zimmerman, 1978).

Since the absolute value of the negative reaction of both high and low ESG-performing firms is not of a high magnitude, inferences of this and the work of others, who reported results of similar magnitude as well, need to be interpreted carefully.

Tables 9 and 10 unroll the relevance of firm size (measured as the natural logarithm of total assets) for ESG performance and IV predicted firm value. That relationship is, for all four coefficients in the lower area flipped compared to the higher area in Table 10. In Table 9, for the first stage of LowESG and LowESGSector, ln(SIZE) is more pronounced in the sectoral partition, whereas size is more pronounced for the first stage in the ESG performance partition in Table 10. In the first stage (ESG performance as dependent variable), firm size seems to be of higher importance for firms in ex-ante lower ESG sectors than for firms with lower ESG performance. In the lower area, sector association seems more critical than prior ESG performance as a predictor of post-directive ESG performance. The relationship is reversed in the first stage for firms with higher ESG performance and those belonging to higher ESG sectors (Table 10). Here, ex-ante sector affiliation plays a minor role compared to ex-ante ESG performance.

Similar to the reversed characteristics for low/high ESG(Sector) firms, the same pattern holds for the second stage of the IV2SLS that was run. In the lower score partition, Tobin's Q is more negatively affected by size for firms with ex-ante lower ESG performance than for firms who belong to ex-ante lower-performing sectors (Table 9). However, ex-ante sector affiliation seems more size-sensitive than prior ESG performance, as findings in in Table 10 suggest.

The relationship of firm size with Tobin's Q as the measure of interest in this work with regards to pre-directive ESG performance and sector affiliation can be summarized into the deduction that the value relevance of size is more pronounced for firms with a HighESGSector affiliation than firms with a prior high level of ESG performance (Table 10). On the contrary, in the lower ex-ante group (performance and sector, Table 9), the value relevance of size is more pronounced for firms with low ESG performance than those in LowESGSectors.

Overall, the results of the two cross-sectional analyses rejected but could explain the results for hypothesis 2a, were able to confirm hypothesis 2b, and further strengthened the

Table 10: Impact of the NFRD on ESG Performance and Tobin's Q - HighESG(Sector)

	HighESG		HighESGSector	
	First Stage (ESG)	Second Stage (Q)	First Stage (ESG)	Second Stage (Q)
Intercept	7.744	13.731***	9.045	20.056***
	(29.660)	(2.215)	(19.096)	(2.737)
coefficient 1 on β or ϑ	2.792**	-0.092***	2.622	-0.099***
	(1.095)	(0.028)	(2.260)	(0.021)
ln(SIZE)	1.761**	-0.204***	1	-0.46***
	(0.898)	(0.066)	(0.612)	(0.092)
LEV	-1.794	1.058*	0.008	0.174
	(3.427)	(0.601)	(4.357)	(0.654)
Firm fixed effects	Included	Included	Included	Included
Year fixed effects	Included	Included	Included	Included
Year-Sector fixed effects	Included	Included	Included	Included
Adjusted R^2	0.791	0.79	0.856	0.817
# of observations	2076	2076	2497	2497

The shown table depicts results of the two stages of the IV2SLS analysis conducted for both partitions to test H2a and H2b, HighESG and HighSectorESG. HighESG includes firms and their matched controls who have been in the upper tercile of ESG performance in the year of matching 2013. HighSectorESG includes firms who are associated with the upper three ESG performance sectors, namely Real Estate, Consumer Non-Cyclicals, and Healthcare ("HighESGSector"). Coefficient 1 on β and ϑ is the main variable of interest. It includes either: the coefficent from equation (6) for the first stage of the IV, β which is the coefficient on the interaction term of two dummy variables, *Treatment* (1 if in treatment group (EU), 0 if in control group (U.S.)) and *Mandate* (1 if after shock date in 2014 (2014-2019), 0 if before shock date (2011-2013)). Or, it includes coefficient ϑ which is the coefficient on the predicted ESG score in the second stage. Dependent variable of interest is ESG performance measured by ESG score in the first stage and Tobin's Q in the second stage. All data retrieved via Refinitiv Eikon Data API. Control variables are: ln(SIZE) and LEV. ln(SIZE) is the natural logarithm of total assets. LEV is the total liabilities over total assets. ***, **, and * indicate statistical significance either at the 1%, 5%, or the 10% level (two-tailed), respectively. Following Flammer (2015) and Gibbons (2020), standard errors are included in parentheses. Standard errors clustered at the firm level are employed for the Low/HighESG partition. Standard errors clustered at the sector level are used for the Low/HighESGSector partition. The model was robust to running all combinations of fixed effects. All variables are defined in Appendix A.

results of hypotheses 1a-c, which were also confirmed earlier. A third cross-sectional analysis regarding the pre-directive reporting status was conducted but discarded and is not reported since results were based on too few firms in the sample that did not report ESG information prior to the directive, which would have made results too unreliable. In the treatment group of 354 EU firms, only 54 firms did not report ESG information in either a standalone report or a section in its annual report in 2013. A potential contribution to an explanation for why there was no substantial firm value effect could be inferred from the low number of firms not reporting prior to the announcement. As there was already a high level of ESG performance maturity, no substantial firm value effect could be detected (Ottenstein et al., 2022). The approach was operationalized with a measure for the pre-directive reporting status. Following De Villiers et al. (2017), Fiechter et al. (2022), Ottenstein et al. (2022), and Stolowy and Paugam (2018), the partition was based on Refinitiv item "CGVSDP026" as a proxy for whether a sample firm reports ESG information in either a standalone report or in a section in its annual report[22] in year t or not.

Based on prior research that showed mainly positive market reactions to mandatory ESG reporting outside the EU (see Chapter 3.1), the findings of this work allow for the potential deduction that either the flawed construction of the NFRD or the information of newly reporting firms caused the negative value reaction in the market. Since only a small number of firms were first-time reporters in the sample, market reaction to this new information must have been considerable, and/or the effect stemmed from the construction of the NFRD itself. As the knowledge of the market reaction is crucial to evaluating the overall effect of the NFRD on firm value, further examination should continue exploring this avenue of research.

6.4. Robustness Tests

To strengthen the above empirical findings, four supplemental robustness checks are included in the following section. First, in the spirit of a recommendation by Larcker and Rusticus (2010) and an exemplary application in Fatemi et al. (2018) and Gibbons (2020), OLS regression results are

[22] This approach was in line with the NFRD requirement that requires a standalone report or an integration in the annual report.

compared to the estimates from the 2SLS analysis. Larcker and Rusticus (2010) advocate that this may mitigate potential difficulties accompanying the IV approach resulting from selecting instrumental variables that meet both the relevance and exogeneity condition for a sound conclusion. The test has been undertaken by contrasting the overall implications from the analysis for answer ***H1a***, which were obtained in an IV setting, with the results from the yearly effects model, which was realised with a two-stage OLS regression (***H1c***). Inferences from this comparison were already mentioned in the *Early and Late Post-Period Firm Value Effects* ***(H1c)*** section. Moreover, equation (3) runs in an OLS setting, which as well resulted in a low negative and significant outcome. Similar results and significance levels suggest that, on average, post-directive ESG performance increased for treatment firms, and the observed market reaction to mandatory ESG disclosure in the EU was weakly negative. This further strengthens the implications of this work.

Second, in accordance with the results of the section *Aggregated ESG* ***(H1a)*** *and Individual E, S, and G* ***(H1b)*** *Firm Value Effects*, in Chapter 6.2 *Baseline Analysis* and the approach of Ioannou and Serafeim (2017), the Governance score is included as a further control variable in both stages of the baseline 2SLS model (equations (2) and (3)) with the Environmental and the Social scores used as dependent variable instead of the overall ESG score. This serves as a simultaneous check which increases statistical resilience since it reduces potential omitted variable bias in this work as Bajic and Yurtoglu (2018) explain. Results in Chapter 6.2 showed that the Governance score was the driving pillar for the control group's combined ESG performance. *Appendix B* shows the individual score development and one plot with the Governance score excluded, where this notion becomes visible. As extant literature and the results of this work demonstrate, each pillar of the ESG score has a different effect, and especially the Governance score has a differentiated effect from the two other scores (Adams et al., 2009; Bajic & Yurtoglu, 2018; Cek & Eyupoglu, 2020; Paolone et al., 2022). Hence, Ioannou and Serafeim's (2017) line of thought is adopted. They argue that "if firms' increase in governance disclosure is associated with improvements in corporate governance and increases in governance disclosure are also associated with increases in environmental and social disclosure, then the results suggesting an association between increases in ESG disclosure and firm value could be the result of improvements in corporate governance." (Ioannou & Serafeim, 2017, p. 26). Results for this robustness test are recorded in Table 11 in *Appendix D*. Overall inferences of an increase in ESG performance and a decrease in firm value remain unchanged. Resultingly, it can be deduced that changes in Governance scores (performance) do not solely influence firm value or ESG scores. Not surprisingly, based on the results of the analysis for ***H1b*** and observable in the plotted development of each subcomponents' performance in *Appendix B*, the Social performance is the main driver for the positive development of the overall ESG performance. However, the Environmental performance seems more relevant for the firm value, as the second stage in Table 11 depicts. This test's importance arises from the potential endogeneity problem associated with corporate governance and firm value, which can be addressed by utilizing the governance performance as a control variable (Ioannou & Serafeim, 2017).

Third, a placebo event is placed in 2012, two years prior to the shock, in line with Atanasov and Black (2016) and the application by Ioannou and Serafeim (2017) and Krueger (2015a). Specifically, the IV2SLS model used to test ***H1a*** was run for the changed shock date and the exact main specifications, and no significant results were obtained. This stands in line with the findings depicted in Figure 2.

The statistical power is additionally elevated as the sample obtained after propensity score matching is relatively large in absolute terms as well as in comparison to other works in the field of research on the implications of the NFRD (e.g., Cuomo et al. (2022), who obtain a sample of 656 firms; Agostini et al. (2022), who work with a sample of 20 firms; Grewal et al. (2019), who receive a sample of 380 firms when additionally matching on ESG performance, and Carnini Pulino et al. (2022), who obtain a sample of 263 firms). All regression models used in this work are run with widely employed model specifications such as utilizing fixed effects (Firm, Year, and Sector-Year), using an IV2SLS setting, and clustering of standard errors. As a fourth robustness test, the three fixed effects are excluded one by one from the baseline model. No diverging effects to the inferences based on the estimated models can be reported (Fatemi et al., 2018). Information on variable correlations are recorded in Table 12 in *Appendix E*.

7. Conclusion

7.1. Summary and Contributions

Revisiting Milton Friedmans' previously quoted view that "The social responsibility of business is to increase its profits" (Friedman, 1970), the findings of this work could potentially fall into his line of reasoning. Robust evidence provided in this work suggests a negative market reaction following the introduction of the NFRD. Nevertheless, this work does not include research on the channels of value destruction. Prior research seems to coincide that higher ESG performance is associated with higher firm value. Thus, one might argue that it is not the market reaction to ESG performance but the enforcement of information disclosure itself or the construction of the directive that creates the negative reaction. It is to be left to future research to explore these drivers.

The implications of the first-time mandate of ESG information disclosure in the EU are examined using an instrumental variable and difference-in-difference approach on a propensity score matched sample. Difference-in-difference results suggest that firm's ESG performance, measured by Refinitiv's ESG scores, significantly increased after the adoption of the directive. Subsequent instrumental variables analysis indicates that the increased ESG performance resulting from the directive is associated with statistically significant

negative effects for Tobin's Q as the measure of firm value. Furthermore, findings indicate that based on the employed sample of 708 firms based in the EU and the U.S., significant anticipatory effects existed. The most considerable effect on ESG performance and Tobin's Q emerges in the first statistically significant year (2015 and 2014, respectively). Moreover, this work finds that sector affiliation and prior ESG performance of firms were pertinent to its valuation in the form that firms with higher ESG ratings or those who belong to sectors with an on average high ESG rating carried a more negative market reaction than firms with lower ex-ante ESG ratings or those belonging to lower rated sectors.

To the best of current understanding, this work offers the first approach to research on firm value implication of the NFRD with an instrumental variables analysis, backed by results from difference-in-difference analyses. Moreover, it is the first work that analyses the implications of the NFRD on Tobin's Q in a yearly effects model. Given the contemporary body of knowledge, this work expands scant research that deals with the value implications of the NFRD and further adds to the notion that the NFRD had an overall negative effect on firm value (Fiechter et al., 2022; Grewal et al., 2019; Mittelbach-Hörmanseder et al., 2021). Furthermore, research on the anticipatory effects of mandatory disclosure regulation (Fiechter et al., 2022; Grewal et al., 2019) as well as the relevance of sectoral affiliation (Cahan et al., 2016; Christensen et al., 2017; Eccles et al., 2012; Krueger, 2015a) and prior ESG performance (Fiechter et al., 2022; Grewal et al., 2019; Ioannou & Serafeim, 2017; Jackson et al., 2020; Mittelbach-Hörmanseder et al., 2021) is extended with the observations in this work. Moreover, the findings further support literature in the field of disclosure and its relationship with ESG/CSR activities (Christensen et al., 2017; Cutler et al., 2004; Schlenker & Scorse, 2017).

More generally, the examination at hand expands earlier scholarly efforts in the field of information disclosure regulation (Armstrong et al., 2010; Bolton et al., 2021; Daske et al., 2008; Leuz & Wysocki, 2015) and can have implications for the field of Mergers & Acquisitions as well (Fich et al., 2021; T. Li et al., 2023). Furthermore, it adds to literature on non-financial disclosure (Cheng et al., 2014; Dhaliwal et al., 2011). Overall, the outcomes of this work may offer insights for further research, policymakers, and practitioners alike.

7.2. Implications for Managerial Practice and Policy

Four central insights for managerial practice and policy are derived following the analysis of the first mandated ESG disclosure regulation in the European Union. First, results suggest that overall, the introduction of the NFRD proved to be a significant event for the information landscape of capital markets, which condensates into an (negative) effect on firm value, as depicted in this work. Hence, it can be deducted that investors view the directive as costly and attribute specific costs and/or benefits to the directive. Second, results of the baseline analysis and the cross-sectional analyses, which were mostly in line with prior research, have revealed that, on average and based on the sample and methods employed, the NFRD had weakly negative firm value implications that differed across firms. From these, it can be inferred that if managers of high ESG-performing firms anticipate a more negative market reaction to a directive than those of low ESG-performing firms, they would be incentivized to disclose less ESG information or reduce their ESG activities to reduce the negative market reaction. Since the goal of the NFRD was to increase the information environment for market participants and spark an internalization of externalities through market mechanisms (see Grewal et al., 2019), such management behaviour would be counterproductive.

Third, because the negative market reaction might partly stem from the flaws in its construction, as the European Union itself had to recognize (Baumüller & Grbenic, 2021; European Parliament, 2022), it does not incentivize firms to disclose or improve ESG performance. Hence, an alteration to the directive (which was made in form of the CSRD) can improve the incentivization and materialize the long-range efforts of the EU. Thus, this work could serve as information for future improvement work on ESG disclosure jurisdiction in general. Fourth, moreover the results of this work have important implications for ESG information disclosure regulation in other jurisdictions, as e.g., the market reaction following the upcoming first-time announcement of climate-related disclosure requirements by the U.S. Securities and Exchange Commission in October 2023 can be predicted more precisely (McGowan, 2023). Additionally, by conducting cross-comparisons of market reactions to the NFRD with other ESG reporting mandates in the EU, such as the CSRD, or outside the EU, such as Integrated Reporting in South Africa, policymakers can draw conclusions about how to construct such a mandate best.

7.3. Limitations and Future Research

The findings of this work are subject to limitations. First, the analysis includes a sample with data obtained from 2011 to 2019. It is pertinent to mention that as the NFRD came into force in 2017, only two years after this date are covered in the sample since the Covid-19 pandemic impacted markets from the beginning of 2020 onwards. The same situation holds for the pre-directive period, where due to various reasons laid out in the temporal considerations section, data was not collected earlier. Hence, this work reports results that a time frame of limited extent could flaw. Second, it is invariably possible that other shocks, which have not been accounted for by the thoroughly conducted analyses, such as the mentioned Paris Climate Agreement or elections in the EU or the U.S., had different effects on each group.

Third, this work draws conclusions on associations, not on causal relationships. Hence, it cannot be ruled out entirely that reverse causality might influence obtained results. Fourth, as models are only as good as their input, this work can be subject to bias or errors caused by the data provided by Refinitiv's Eikon Data API. Specifically, as current research shows, ESG ratings can diverge by rating agency and should be treated with caution because of, e.g., rater effects, where

a rater's total view of a company affects the measurement of particular ESG categories (Berg et al., 2022). Consequently, further research should consult more than one data provider, conduct manual data collection, invest in developing category-specific metrics, and bear in mind that reported data may be structurally flawed by incentivization. Fifth, it cannot be conclusively negated that greenwashing might partly influence the findings of this work. The measure of ESG scores, as a proxy for ESG performance, could potentially include merely symbolic activities that Refinitiv has not detected.

Henceforward, as not analysed in this work, further avenues for research in the contemporary field of mandatory ESG disclosure and firm value implications of the latter may include research on the channels of value creation. To not only draw out the overall macroeconomic value implication of the directive but also to illustrate a more nuanced palette of implications with deeper managerial insights, further research, especially into the channels of value creation or destruction, need to be conducted (e.g., as mentioned, costs, benefits, good/bad ESG information valuation, the legal enforcement of the NFRD and the "comply-or-explain-approach"). Academic research could additionally address the underlying mechanisms by studying anecdotal evidence with individual firm reactions. Furthermore, scholars should leverage the conceptual opportunity to exploit the enlarged potential sample resulting from the individual national transposition. As mentioned, this increased the number from around 2,000 to an estimated 11,500 firms in the EU that had to comply with the NFRD. Simultaneously, studies could employ other measures of firm value, especially different versions of Tobin's Q, as research suggests (see Bartlett and Partnoy, 2018).

References

Adams, R. B., Hermalin, B. E., & Weisbach, M. S. (2009). The Role of Boards of Directors in Corporate Governance: A Conceptual Framework & Survey. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1299212

Agliati, M. M. (2021). An Economic Analysis and Review of Non-Financial Reporting in the European Union.

Agostini, M., Costa, E., & Korca, B. (2022). Non-financial disclosure and corporate financial performance under directive 2014/95/EU: Evidence from Italian listed companies. *Accounting in Europe*, *19*(1), 78–109. https://doi.org/10.1080/17449480.2021.1979610

Ahmad, M. F., Aktas, N., & Croci, E. (2023). Climate risk and deployment of corporate resources to working capital. *Economics Letters*, *224*, 111002. https://doi.org/10.1016/j.econlet.2023.111002

Allman, E., & Won, J. (2022). The Effect of ESG Disclosure on Corporate Investment Efficiency. *Proceedings of Paris December 2021 Finance Meeting EUROFIDAI - ESSEC*. https://doi.org/10.2139/ssrn.3816592

Amel-Zadeh, A., & Serafeim, G. (2017). Why and How Investors Use ESG Information: Evidence from a Global Survey. *Financial Analysts Journal*, *74*(3), 87–103.

Amihud, Y., & Mendelson, H. (1986). Asset Pricing and the Bid-Ask Spread. *Journal of Financial Economics*, *17*(2), 223–249. https://doi.org/10.1016/0304-405X(86)90065-6

Angrist, J. D., & Krueger, A. B. (2001). Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments. *Journal of Economic Perspectives*, *15*(4), 69–85. https://doi.org/10.1257/jep.15.4.69

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Armstrong, C. S., Barth, M. E., Jagolinzer, A. D., & Riedl, E. J. (2010). Market Reaction to the Adoption of IFRS in Europe. *The Accounting Review*, *85*(1), 31–61. https://doi.org/10.2308/accr.2010.85.1.31

Atanasov, V., & Black, B. (2016). Shock-Based Causal Inference in Corporate Finance and Accounting Research. *Critical Finance Review*, *5*(2), 207–304. https://doi.org/10.1561/104.00000036

Bajic, S., & Yurtoglu, B. (2018). Which aspects of CSR predict firm market value? *Journal of Capital Markets Studies*, *2*(1), 50–69. https://doi.org/10.1108/jcms-10-2017-0002

Baran, L. C., & King, T. H. D. (2014). S&P 500 Index reconstitutions and information asymmetry. *Applied Financial Economics*, *24*(11), 777–791. https://doi.org/10.1080/09603107.2014.904489

Barth, M. E., Cahan, S. F., Chen, L., & Venter, E. R. (2017). The economic consequences associated with integrated report quality: Capital market and real effects. *Accounting, Organizations and Society*, *62*, 43–64. https://doi.org/10.1016/j.aos.2017.08.005

Bartlett, R. P., & Partnoy, F. (2018). The Misuse of Tobin's Q. https://doi.org/10.2139/ssrn.3118020

Bascle, G. (2008). Controlling for endogeneity with instrumental variables in strategic management research. *Strategic Organization*, *6*(3), 285–327. https://doi.org/10.1177/1476127008094339

Baumüller, J., & Grbenic, S. O. (2021). MOVING FROM NON-FINANCIAL TO SUSTAINABILITY REPORTING: ANALYZING THE EU COMMISSION'S PROPOSAL FOR A CORPORATE SUSTAINABILITY REPORTING DIRECTIVE (CSRD). *Facta Universitatis, Series: Economics and Organization*, *1*, 369. https://doi.org/10.22190/FUEO210817026B

Berg, F., Kölbel, J. F., & Rigobon, R. (2022). Aggregate Confusion: The Divergence of ESG Ratings. *Review of Finance*, *26*(6), 1315–1344. https://doi.org/10.1093/rof/rfac033

Bhagawan, P., & Mukhopadhyay, J. P. (2019). Does Mandatory Expenditure on CSR Affect Firm Value? Evidence from Indian Firms. *2019 Financial Markets & Corporate Governance Conference*. https://ssrn.com/abstract=3319214

Bolton, P., Kacperczyk, M. T., Leuz, C., Ormazabal, G., Reichelstein, S., & Schoenmaker, D. (2021). Mandatory Corporate Carbon Disclosures and the Path to Net Zero. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3946031

Buallay, A. (2019). Between cost and value: Investigating the effects of sustainability reporting on a firm's performance. *Journal of Applied Accounting Research*, *20*(4), 481–496. https://doi.org/10.1108/JAAR-12-2017-0137

Cahan, S. F., De Villiers, C., Jeter, D. C., Naiker, V., & Van Staden, C. J. (2016). Are CSR Disclosures Value Relevant? Cross-Country Evidence. *European Accounting Review*, *25*(3), 579–611. https://doi.org/10.1080/09638180.2015.1064009

Carnini Pulino, S., Ciaburri, M., Magnanelli, B. S., & Nasta, L. (2022). Does ESG Disclosure Influence Firm Performance? *Sustainability*, *14*(13), 7595. https://doi.org/10.3390/su14137595

Cek, K., & Eyupoglu, S. (2020). Does Environmental, Social and Governance Performance Influence Economic Performance? *Journal of Business Economics and Management*, *21*(4), 1165–1184. https://doi.org/10.3846/jbem.2020.12725

Chang, X., Fu, K., Jin, Y., & Liem, P. F. (2022). Sustainable Finance: ESG/CSR, Firm Value, and Investment Returns. *Asia-Pacific Journal of Financial Studies*, *51*(3), 325–371. https://doi.org/10.1111/ajfs.12379

Chen, Y.-C., Hung, M., & Wang, Y. (2018). The effect of mandatory CSR disclosure on firm profitability and social externalities: Evidence from China. *Journal of Accounting and Economics*, *65*(1), 169–190. https://doi.org/10.1016/j.jacceco.2017.11.009

Cheng, B., Ioannou, I., & Serafeim, G. (2014). Corporate social responsibility and access to finance: CSR and Access to Finance. *Strategic*

Management Journal, *35*(1), 1–23. https://doi.org/10.1002/smj.2131

Cho, C. H., Laine, M., Roberts, R. W., & Rodrigue, M. (2015). Organized hypocrisy, organizational façades, and sustainability reporting. *Accounting, Organizations and Society*, *40*, 78–94. https://doi.org/10.1016/j.aos.2014.12.003

Cho, C. H., Michelon, G., Patten, D. M., & Roberts, R. W. (2015). CSR disclosure: The more things change...? *Accounting, Auditing & Accountability Journal*, *28*(1), 14–35. https://doi.org/10.1108/AAAJ-12-2013-1549

Choi, D., Gao, Z., & Jiang, W. (2020). Attention to Global Warming. *The Review of Financial Studies*, *33*(3), 1112–1145. https://doi.org/10.1093/rfs/hhz086

Christensen, H. B., Floyd, E., Liu, L. Y., & Maffett, M. (2017). The real effects of mandated information on social responsibility in financial reports: Evidence from mine-safety records. *Journal of Accounting and Economics*, *64*(2–3), 284–304. https://doi.org/10.1016/j.jacceco.2017.08.001

Christensen, H. B., Hail, L., & Leuz, C. (2018). Economic Analysis of Widespread Adoption of CSR and Sustainability Reporting Standards. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3315673

Christensen, H. B., Hail, L., & Leuz, C. (2021). Mandatory CSR and sustainability reporting: Economic analysis and literature review. *Review of Accounting Studies*, *26*(3), 1176–1248. https://doi.org/10.1007/s11142-021-09609-5

Cicchiello, A. F., Marrazza, F., & Perdichizzi, S. (2023). Non-financial disclosure regulation and environmental, social, and governance (ESG) performance: The case of EU and US firms. *Corporate Social Responsibility and Environmental Management*, *30*(3), 1121–1128. https://doi.org/10.1002/csr.2408

Cifrino, D. A. (2023). The Rise of International ESG Disclosure Standards. https://corpgov.law.harvard.edu/2023/06/29/the-rise-of-international-esg-disclosure-standards/

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330–351. https://doi.org/10.1037/1082-989X.6.4.330

Constantinides, G. M. (1986). Capital Market Equilibrium with Transaction Costs. *Journal of Political Economy*, *94*(4), 842–862. http://dx.doi.org/10.1086/261410

Conway, E. (2019). Quantitative impacts of mandatory integrated reporting. *Journal of Financial Reporting and Accounting*, *17*(4), 604–634. https://doi.org/10.1108/JFRA-08-2018-0066

Costa, E., & Agostini, M. (2016). Mandatory Disclosure about Environmental and Employee Matters in the Reports of Italian-Listed Corporate Groups. *Social and Environmental Accountability Journal*, *36*(1), 10–33. https://doi.org/10.1080/0969160X.2016.1144519

Crifo, P., Forget, V. D., & Teyssier, S. (2015). The price of environmental, social and governance practice disclosure: An experiment with professional private equity investors. *Journal of Corporate Finance*, *30*, 168–194. https://doi.org/10.1016/j.jcorpfin.2014.12.006

Cuomo, F., Gaia, S., Girardone, C., & Piserà, S. (2022). The effects of the EU non-financial reporting directive on corporate social responsibility. *The European Journal of Finance*, 1–27. https://doi.org/10.1080/1351847X.2022.2113812

Cutler, D. M., Huckman, R. S., & Landrum, M. B. (2004). The Role of Information in Medical Markets: An Analysis of Publicly Reported Outcomes in Cardiac Surgery. *American Economic Review*, *94*(2), 342–346. https://doi.org/10.1257/0002828041301993

D'Amato, A., & Falivena, C. (2020). Corporate social responsibility and firm value: Do firm size and age matter? Empirical evidence from European listed companies. *Corporate Social Responsibility and Environmental Management*, *27*(2), 909–924. https://doi.org/10.1002/csr.1855

Darendeli, A., Fiechter, P., Hitz, J.-M., & Lehmann, N. (2022). The role of corporate social responsibility (CSR) information in supply-chain contracting: Evidence from the expansion of CSR rating coverage. *Journal of Accounting and Economics*, *74*(2–3), 101525. https://doi.org/10.1016/j.jacceco.2022.101525

Daske, H., Hail, L., Leuz, C., & Verdi, R. S. (2008). Mandatory IFRS Reporting Around the World: Early Evidence on the Economic Consequences. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1024240

Davies, P. A., Little, D., Green, M. D., Bee, J., & Pierce, A. J. (2023). The EU Corporate Sustainability Reporting Directive—How Companies Need to Prepare. https://www.lw.com/admin/upload/SiteAttachments/Alert%203059.pdf

De Andrés, P., De La Fuente, G., & Velasco, P. (2017). Does it really matter how a firm diversifies? Assets-in-place diversification versus growth options diversification. *Journal of Corporate Finance*, *43*, 316–339. https://doi.org/10.1016/j.jcorpfin.2017.01.011

De Groen, W. P., Alcidi, C., Simonelli, F., Campmas, A., Di Salvo, M., Musmeci, R., Oliinyk, I., & Tadi, S. (2020). *Study on the non-financial reporting directive: Final report* (tech. rep.). https://data.europa.eu/doi/10.2874/229601

De Villiers, C., Venter, E. R., & Hsiao, P.-C. K. (2017). Integrated reporting: Background, measurement issues, approaches and an agenda for future research. *Accounting & Finance*, *57*(4), 937–959. https://doi.org/10.1111/acfi.12246

Deloitte Touche Tohmatsu Limited. (2023). Global Reach of the E.U. Corporate Sustainability Reporting Directive and the Impact on U.S. Companies. https://dart.deloitte.com/USDART/home/publications/deloitte/heads-up/2023/esg-eu-corporate-sustainability-reporting

Deng, X., Kang, J., & Low, B. S. (2013). Corporate social responsibility and stakeholder value maximization: Evidence from mergers. *Journal of Financial Economics*, *110*(1), 87–109. https://doi.org/10.1016/j.jfineco.2013.04.014

Dhaliwal, D. S., Li, O. Z., Tsang, A., & Yang, Y. G. (2011). Voluntary Nonfinancial Disclosure and the Cost of Equity Capital: The Initiation of Corporate Social Responsibility Reporting. *The Accounting Review*, *86*(1), 59–100. https://doi.org/10.2308/accr.00000005

Drempetic, S., Klein, C., & Zwergel, B. (2020). The Influence of Firm Size on the ESG Score: Corporate Sustainability Ratings Under Review. *Journal of Business Ethics*, *167*(2), 333–360. https://doi.org/10.1007/s10551-019-04164-1

Easley, D., & O'Hara, M. (2004). Information and the Cost of Capital. *The Journal of Finance*, *59*(4), 1553–1583.

Eccles, R. G., Ioannou, I., & Serafeim, G. (2014). The Impact of a Corporate Culture of Sustainability on Corporate Behavior and Performance. *Management Science*, *60*(11), 2835–2857. https://doi.org/10.2139/ssrn.1964011

Eccles, R. G., Krzus, M. P., Rogers, J., & Serafeim, G. (2012). The Need for Sector-Specific Materiality and Sustainability Reporting Standards. *Journal of Applied Corporate Finance*, *24*(2), 65–71. https://doi.org/10.1111/j.1745-6622.2012.00380.x

Eccles, R. G., Serafeim, G., & Krzus, M. P. (2011). Market Interest in Nonfinancial Information. *Journal of Applied Corporate Finance*, *23*(4), 113–127. https://doi.org/10.1111/j.1745-6622.2011.00357.x

European Commission. (2013). DIRECTIVE 2013/34/EU OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 26 June 2013 on the annual financial statements, consolidated financial statements and related reports of certain types of undertakings, amending Directive 2006/43/EC of the European Parliament and of the Council and repealing Council Directives 78/660/EEC and 83/349/EEC.

European Commission. (2014a). DIRECTIVE 2014/95/EU OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 22 October 2014 amending Directive 2013/34/EU as regards disclosure of non-financial and diversity information by certain large undertakings and groups. https://doi.org/10.4337/9781800372092.00015

European Commission. (2014b). Disclosure of non-financial and diversity information by large companies and groups—Frequently asked questions. https://ec.europa.eu/commission/presscorner/detail/cs/MEMO_14_301

European Commission. (2019a). REGULATION (EU) 2019/2088 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 November 2019 on sustainability-related disclosures in the financial services sector. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32019R2088

European Commission. (2019b). REGULATION (EU) 2019/876 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 20 May 2019 amending Regulation (EU) No 575/2013 as regards the leverage ratio, the net stable funding ratio, requirements for own funds and eligible liabilities, counterparty credit risk, market risk, exposures to central counterparties, exposures to collective investment undertakings, large exposures, reporting and disclosure requirements, and Regulation (EU) No 648/2012. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32019R0876

European Commission. (2019c). The European Green Deal sets out how to make Europe the first climate-neutral continent by 2050, boosting the economy, improving people's health and quality of life, caring for nature, and leaving no one behind. https://ec.europa.eu/commission/presscorner/detail/en/ip_19_6691

European Commission. (2020). REGULATION (EU) 2020/852 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 18 June 2020 on the establishment of a framework to facilitate sustainable investment, and amending Regulation (EU) 2019/2088. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32020R0852&from=EN

European Commission. (2022). DIRECTIVE (EU) 2022/2464 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 14 December 2022 amending Regulation (EU) No 537/2014, Directive 2004/109/EC, Directive 2006/43/EC and Directive 2013/34/EU, as regards corporate sustainability reporting. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022L2464

European Parliament. (2022). Sustainable economy: Parliament adopts new reporting rules for multinationals. https://www.europarl.europa.eu/pdfs/news/expert/2022/11/press_release/20221107IPR49611/20221107IPR49611_en.pdf

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance*, *25*(2), 383. https://doi.org/10.2307/2325486

Fama, E. F., & French, K. R. (1992). The Cross-Section of Expected Stock Returns. *The Journal of Finance*, *47*(2), 427–465. https://doi.org/10.1111/j.1540-6261.1992.tb04398.x

Fama, E. F., & French, K. R. (2002). Testing Trade-Off and Pecking Order Predictions About Dividends and Debt. *Review of Financial Studies*, *15*(1), 1–33. https://doi.org/10.1093/rfs/15.1.1

Fatemi, A., Glaum, M., & Kaiser, S. (2018). ESG performance and firm value: The moderating role of disclosure. *Global Finance Journal*, *38*, 45–64. https://doi.org/10.1016/j.gfj.2017.03.001

Fiandrino, S., Gromis Di Trana, M., Tonelli, A., & Lucchese, A. (2022). The multi-faceted dimensions for the disclosure quality of non-financial information in revising directive 2014/95/EU. *Journal of Applied Accounting Research*, *23*(1), 274–300. https://doi.org/10.1108/JAAR-04-2021-0118

Fich, E. M., Griffin, T., & Kalmenovitz, J. (2021). Does Regulatory Exposure Create M&A Synergies? https://doi.org/10.2139/ssrn.3966359

Fiechter, P., Hitz, J., & Lehmann, N. (2022). Real Effects of a Widespread CSR Reporting Mandate: Evidence from the European Union's CSR Directive. *Journal of Accounting Research*, *60*(4), 1499–1549. https://doi.org/10.1111/1475-679X.12424

Flammer, C. (2013). Corporate Social Responsibility and Shareholder Reaction: The Environmental Awareness of Investors. *Academy of Management Journal*, *56*(3), 758–781. https://doi.org/10.5465/amj.2011.0744

Flammer, C. (2015). Does Corporate Social Responsibility Lead to Superior Financial Performance? A Regression Discontinuity Approach. *Management Science*, *61*(11), 2549–2568. https://doi.org/10.1287/mnsc.2014.2038

Flammer, C. (2021). Corporate green bonds. *Journal of Financial Economics*, *142*(2), 499–516. https://doi.org/10.1016/j.jfineco.2021.01.010

Freeman, R. E., & Dmytriyev, S. (2017). Corporate Social Responsibility and Stakeholder Theory: Learning From Each Other. *Symphonya. Emerging Issues in Management*, *1*, 7–15. https://doi.org/10.4468/2017.1.02freeman.dmytriyev

Frieden, J., & Walter, S. (2017). Understanding the Political Economy of the Eurozone Crisis. *Annual Review of Political Science*, *20*(1), 371–390. https://doi.org/10.1146/annurev-polisci-051215-023101

Friedman, M. (1970). The Social Responsibility of Business Is to Increase Its Profits. *New York Times Magazine*, 122–126.

Fuente, G. D. L., Ortiz, M., & Velasco, P. (2022). The value of a firm's engagement in ESG practices: Are we looking at the right side? *Long Range Planning*, *55*(4), 102143. https://doi.org/10.1016/j.lrp.2021.102143

G&A. (2021). Flash Report S&P 500 2020—Trends on the sustainability reporting practices of S&P 500 Index companies. *Governance & Accountability Institute, Inc.*

G&A. (2023). All-Time High of Sustainability Reports Among U.S. Publicly-Traded Companies: 96% of S&P 500 and 81% of Russell 1000. https://www.ga-institute.com/research/ga-research-directory/sustainability-reporting-trends/2022-sustainability-reporting-in-focus.html

Gibbons, B. (2020). The Financially Material Effects of Mandatory Non-Financial Disclosure [Forthcoming]. *Journal of Accounting Research*. http://dx.doi.org/10.2139/ssrn.3658415

Gibson, R., & Krueger, P. (2017). The Sustainability Footprint of Institutional Investors. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2918926

Google Trends. (2023). Comparison of Google Search Trends between 'ESG' and 'CSR' for the Past Five Years. https://trends.google.com/trends/explore?date=today%205-y&q=esg,csr&hl=de

Gormley, T. A., & Matsa, D. A. (2012). Common Errors: How to (and Not to) Control for Unobserved Heterogeneity. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2023868

Graham, J. W. (2003). Adding Missing-Data-Relevant Variables to FIML-Based Structural Equation Models. *Structural Equation Modeling: A Multidisciplinary Journal*, *10*(1), 80–100. https://doi.org/10.1207/S15328007SEM1001_4

Grewal, J., Riedl, E. J., & Serafeim, G. (2019). Market Reaction to Mandatory Nonfinancial Disclosure. *Management Science*, *65*(7), 3061–3084. https://doi.org/10.1287/mnsc.2018.3099

GRI. (2022). The GRI Standards—Enabling transparency on organizational impacts. https://www.globalreporting.org/media/wmxlklns/about-gri-brochure-2022.pdf

Gross, J., & Zahner, J. (2021). What is on the ECB's mind? Monetary policy before and after the global financial crisis. *Journal of Macroeconomics*, *68*, 103292. https://doi.org/10.1016/j.jmacro.2021.103292

Gulenko, M. (2018). Mandatory CSR reporting—Literature review and future developments in Germany. *NachhaltigkeitsManagementForum | Sustainability Management Forum*, *26*(1–4), 3–17. https://doi.org/10.1007/s00550-018-0476-9

Harper Ho, V. E. (2016). Risk-Related Activism: The Business Case for Monitoring Nonfinancial Risk. *Journal of Corporation Law*, *41*, 647–704. https://doi.org/10.2139/ssrn.2478121

Healy, P. M., & Palepu, K. G. (2001). Information asymmetry, corporate disclosure, and the capital markets: A review of the empirical disclosure literature. *Journal of Accounting and Economics*, *31*(1–3), 405–440. https://doi.org/10.1016/S0165-4101(01)00018-0

Howarth, R. B., & Norgaard, R. B. (1992). Environmental Valuation under Sustainable Development. *The American Economic Review*, *82*(2), 473–477.

Hummel, K., & Jobst, D. (2022). The Current State of Corporate Sustainability Reporting Regulation in the European Union. https://doi.org/10.2139/ssrn.3978478

Huntington-Klein, N. (2021). *The Effect: An Introduction to Research Design and Causality* (1st). Chapman; Hall/CRC. https://doi.org/10.1201/9781003226055

IMF. (2013). *Current Developments in Monetary and Financial Law, Volume 6*. INTERNATIONAL MONETARY FUND. https://doi.org/10.5089/9781616350819.072

Ioannou, I., & Serafeim, G. (2017). The Consequences of Mandatory Corporate Sustainability Reporting. *Harvard Business School Research Working Papers*, (11-100). https://ssrn.com/abstract=1799589

Jackson, G., Bartosch, J., Avetisyan, E., Kinderman, D., & Knudsen, J. S. (2020). Mandatory Non-financial Disclosure and Its Influence on

CSR: An International Comparison. *Journal of Business Ethics*, *162*(2), 323–342. https://doi.org/10.1007/s10551-019-04200-0

Jadiyappa, N., Iyer, S. R., & Jyothi, P. (2021). Does social responsibility improve firm value? Evidence from mandatory corporate social responsibility regulations in India. *International Review of Finance*, *21*(2), 653–660. https://doi.org/10.1111/irfi.12282

Jensen, M. C., & Meckling, W. H. (1978). Can the Corporation Survive? *Financial Analysts Journal*, *31*(1), 31–37. https://doi.org/10.2139/ssrn.244155

Jessop, S., & Kelly, J. (2014). New EU law to help investors pick good corporate citizens. https://www.reuters.com/article/uk-investment-ethical-idUKBREA3K0DF20140421

Jiao, Y. (2010). Stakeholder welfare and firm value. *Journal of Banking & Finance*, *34*(10), 2549–2561. https://doi.org/10.1016/j.jbankfin.2010.04.013

Kim, E.-H., & Lyon, T. P. (2015). Greenwash vs. Brownwash: Exaggeration and Undue Modesty in Corporate Sustainability Disclosure. *Organization Science*, *26*(3), 705–723. https://doi.org/10.1287/orsc.2014.0949

Kitzmueller, M., & Shimshack, J. (2012). Economic Perspectives on Corporate Social Responsibility. *Journal of Economic Literature*, *50*(1), 51–84. https://doi.org/10.1257/jel.50.1.51

Kline, A., & Luo, Y. (2022). PsmPy: A Package for Retrospective Cohort Matching in Python. *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 1354–1357. https://doi.org/10.1109/EMBC48229.2022.9871333

Kolk, A. (2003). Trends in sustainability reporting by the Fortune Global 250. *Business Strategy and the Environment*, *12*(5), 279–291. https://doi.org/10.1002/bse.370

Konar, S., & Cohen, M. A. (1997). Information As Regulation: The Effect of Community Right to Know Laws on Toxic Emissions. *Journal of Environmental Economics and Management*, *32*(1), 109–124. https://doi.org/10.1006/jeem.1996.0955

Krueger, P. (2015a). Climate Change and Firm Valuation: Evidence from a Quasi-Natural Experiment. *Swiss Finance Institute Research Paper*, 15–40. https://doi.org/10.2139/ssrn.2565523

Krueger, P. (2015b). Corporate goodness and shareholder wealth. *Journal of Financial Economics*, *115*(2), 304–329. https://doi.org/10.1016/j.jfineco.2014.09.008

Krueger, P., Sautner, Z., & Starks, L. T. (2020). The Importance of Climate Risks for Institutional Investors. *The Review of Financial Studies*, *33*(3), 1067–1111. https://doi.org/10.1093/rfs/hhz137

Krueger, P., Sautner, Z., Tang, D. Y., & Zhong, R. (2021). The Effects of Mandatory ESG Disclosure around the World. *European Corporate Governance Institute – Finance Working Paper*, (754/2021). https://doi.org/10.2139/ssrn.3832745

La Porta, R., Lopez-de-Silanes, F., Shleifer, A., & Vishny, R. (2000). Investor protection and corporate governance. *Journal of Financial Economics*, *58*(1–2), 3–27. https://doi.org/10.1016/S0304-405X(00)00065-9

La Torre, M., Sabelfeld, S., Blomkvist, M., Tarquinio, L., & Dumay, J. (2018). Harmonising non-financial reporting regulation in Europe: Practical forces and projections for future research. *Meditari Accountancy Research*, *26*(4), 598–621. https://doi.org/10.1108/MEDAR-02-2018-0290

Lambert, R., Leuz, C., & Verrecchia, R. E. (2007). Accounting Information, Disclosure, and the Cost of Capital. *Journal of Accounting Research*, *45*(2), 385–420. https://doi.org/10.1111/j.1475-679X.2007.00238.x

Lambert, R., Leuz, C., & Verrecchia, R. E. (2012). Information Asymmetry, Information Precision, and the Cost of Capital. *Review of Finance*, *16*(1), 1–29. https://doi.org/10.1093/rof/rfr014

Lang, L., Ofek, E., & Stulz, R. M. (1996). Leverage, investment, and firm growth. *Journal of Financial Economics*, *40*(1), 3–29. https://doi.org/10.1016/0304-405X(95)00842-3

Lang, M., Lins, K. V., & Maffett, M. (2012). Transparency, Liquidity, and Valuation: International Evidence on When Transparency Matters Most. *Journal of Accounting Research*, *50*(3), 729–774. https://doi.org/10.1111/j.1475-679X.2012.00442.x

Lang, M., & Lundholm, R. (1993). Cross-Sectional Determinants of Analyst Ratings of Corporate Disclosures. *Journal of Accounting Research*, *31*(2), 246. https://doi.org/10.2307/2491273

Larcker, D. F., & Rusticus, T. O. (2010). On the use of instrumental variables in accounting research. *Journal of Accounting and Economics*, *49*(3), 186–205. https://doi.org/10.1016/j.jacceco.2009.11.004

Lee, K.-W., & Yeo, G. H.-H. (2016). The association between integrated reporting and firm valuation. *Review of Quantitative Finance and Accounting*, *47*(4), 1221–1250. https://doi.org/10.1007/s11156-015-0536-y

Leftwich, R. W., Watts, R. L., & Zimmerman, J. L. (1981). Voluntary Corporate Disclosure: The Case of Interim Reporting. *Journal of Accounting Research*, *19*, 50. https://doi.org/10.2307/2490984

Leuz, C., & Wysocki, P. D. (2008). Economic Consequences of Financial Reporting and Disclosure Regulation: A Review and Suggestions for Future Research. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1105398

Leuz, C., & Wysocki, P. D. (2015). The Economics of Disclosure and Financial Reporting Regulation: Evidence and Suggestions for Future Research. *European Corporate Governance Institute (ECGI) - Law Working Paper*, (306/2016). https://doi.org/10.2139/ssrn.2733831

Li, F., & Thomas, L. E. (2018). Addressing Extreme Propensity Scores via the Overlap Weights. *American Journal of Epidemiology*. https://doi.org/10.1093/aje/kwy201

Li, T., Peng, Q., & Yu, L. (2023). ESG Considerations in Acquisitions and Divestitures: Corporate Responses to Mandatory ESG Disclosure. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4376676

Li, Z., & Jia, J. (2022). Effect of mandatory sustainability disclosure announcements: cross-country evidence. *Pacific Accounting Review*, *34*(1), 127–155. https://doi.org/10.1108/PAR-09-2020-0141

Liang, H., & Renneboog, L. (2020). Corporate Social Responsibility and Sustainable Finance: A Review of the Literature. *European Corporate Governance Institute – Finance Working Paper*, (701/2020). https://doi.org/10.2139/ssrn.3698631

Magill, M., Quinzii, M., & Rochet, J.-C. (2015). A Theory of the Stakeholder Corporation. *Econometrica*, *83*(5), 1685–1725. https://doi.org/10.3982/ECTA11455

Manchiraju, H., & Rajgopal, S. (2017). Does Corporate Social Responsibility (CSR) Create Shareholder Value? Evidence from the Indian Companies Act 2013: IMPACT OF CSR ON SHAREHOLDER VALUE. *Journal of Accounting Research*, *55*(5), 1257–1300. https://doi.org/10.1111/1475-679X.12174

McGarry, C., Connellan, C., MacLennan, J., Low, L., Grazebook, W., De Catelle, W., & Moutia-Bloom, J. (2022). Corporate Sustainability Reporting: New EU rules for large companies and listed SMEs. https://www.whitecase.com/insight-alert/corporate-sustainability-reporting-new-eu-rules-large-companies-and-listed-smes

McGowan, J. (2023). The SEC May Be Overstepping Its Authority In ESG/Climate Related Disclosure Standards. https://www.forbes.com/sites/jonmcgowan/2023/08/30/the-sec-may-be-overstepping-its-authority-in-esg-climate-related-disclosure-standards/?sh=60f475eb1a2f

Mittelbach-Hörmanseder, S., Hummel, K., & Rammerstorfer, M. (2021). The information content of corporate social responsibility disclosure in Europe: An institutional perspective. *European Accounting Review*, *30*(2), 309–348. https://doi.org/10.1080/09638180.2020.1763818

Oreskes, M. (1985). PROTESTS AT COLUMBIA: STUDENTS AND ISSUES HAVE CHANGED SINCE THE 60'S. *The New York Times*, 1, 25.

Ottenstein, P., Erben, S., Jost, S., Weuster, C. W., & Zülch, H. (2022). From voluntarism to regulation: Effects of Directive 2014/95/EU on sustainability reporting in the EU. *Journal of Applied Accounting Research*, *23*(1), 55–98. https://doi.org/10.1108/JAAR-03-2021-0075

Paolone, F., Cucari, N., Wu, J., & Tiscini, R. (2022). How do ESG pillars impact firms' marketing performance? A configurational analysis in the pharmaceutical sector. *Journal of Business & Industrial Marketing*, *37*(8), 1594–1606. https://doi.org/10.1108/JBIM-07-2020-0356

Park, J. G., Park, K., Noh, H., & Kim, Y. G. (2023). Characterization of CSR, ESG, and Corporate Citizenship through a Text Mining-Based Review of Literature. *Sustainability*, *15*(5), 3892. https://doi.org/10.3390/su15053892

Patten, D. M. (1992). Intra-industry environmental disclosures in response to the Alaskan oil spill: A note on legitimacy theory. *Accounting, Organizations and Society*, *17*(5), 471–475. https://doi.org/10.1016/0361-3682(92)90042-Q

Pestana Pavan, P. C. (2020). *Institutional Investor Ownership and the firm value effect: Evidence from Brazil* [Doctoral dissertation, FUNDAÇÃO GETULIO VARGAS ESCOLA DE ADMINISTRAÇÃO DE EMPRESAS DE SÃO PAULO]. https://bibliotecadigital.fgv.br/dspace/bitstream/handle/10438/28886/20200311_Pedro_Cesar_Pestana_Pavan_versao_final%202.pdf

Petersen, M. A. (2009). Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches. *Review of Financial Studies*, *22*(1), 435–480. https://doi.org/10.1093/rfs/hhn053

Pollman, E. (2021). Corporate Social Responsibility, ESG, and Compliance. In *The Cambridge Handbook of Compliance* (pp. 662–672). https://doi.org/doi:10.1017/9781108759458.045

Principles for Responsible Investment. (2022). 2021-22 ANNUAL REPORT. https://www.unpri.org/annual-report-2022

Refinitiv. (2022). ENVIRONMENTAL, SOCIAL AND GOVERNANCE SCORES FROM REFINITIV - May 2022. https://www.refinitiv.com/content/dam/marketing/en_us/documents/methodology/refinitiv-esg-scores-methodology.pdf

Rossi, F., & Harjoto, M. A. (2020). Corporate non-financial disclosure, firm value, risk, and agency costs: evidence from Italian listed companies. *Review of Managerial Science*, *14*(5), 1149–1181. https://doi.org/10.1007/s11846-019-00358-z

Rubin, D. B. (2001). Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation. *Health Services and Outcomes Research Methodology*, *2*(3/4), 169–188. https://doi.org/10.1023/A:1020363010465

Sagayam, S., Ising, E., Leiper-Jennings, S., Leong, V., & Butcher, R. (2022). EUROPEAN UNION'S CORPORATE SUSTAINABILITY REPORTING DIRECTIVE — WHAT NON-EU COMPANIES WITH OPERATIONS IN THE EU NEED TO KNOW. https://www.gibsondunn.com/wp-content/uploads/2022/11/european-union-corporate-sustainability-reporting-directive-what-non-eu-companies-with-operations-in-the-eu-need-to-know.pdf

Schlenker, W., & Scorse, J. (2017). Does Being a "Top 10" Worst Polluter Affect Environmental Releases? Evidence from the U.S. Toxic Release Inventory. *Working Paper, Monterey Institute of International Studies*.

Seaman, S., Galati, J., Jackson, D., & Carlin, J. (2013). What Is Meant by "Missing at Random"? *Statistical Science*, *28*(2). https://doi.org/10.1214/13-STS415

Sheehy, B., & Farneti, F. (2021). Corporate Social Responsibility, Sustainability, Sustainable Development and Corporate Sustainability: What Is the Difference, and Does It Matter? *Sustainability*, *13*(11), 5965. https://doi.org/10.3390/su13115965

Shipman, J. E., Swanquist, Q. T., & Whited, R. L. (2017). Propensity Score Matching in Accounting Research. *The Accounting Review*, *92*(1), 213–244. https://doi.org/10.2308/accr-51449

Sorkin, A. R. (2020). A Free Market Manifesto That Changed the World, Reconsidered. https://www.nytimes.com/2020/09/11/business/dealbook/milton-friedman-doctrine-social-responsibility-of-business.html

S&P Global. (2023). Getting ready for the SEC Climate Disclosure Rule. https://www.spglobal.com/esg/solutions/getting-ready-for-the-sec-climate-disclosure-rule

Stolowy, H., & Paugam, L. (2018). The expansion of non-financial reporting: An exploratory study. *Accounting and Business Research*, *48*(5), 525–548. https://doi.org/10.1080/00014788.2018.1470141

Stroebel, J., & Wurgler, J. (2021). What do you think about climate finance? *Journal of Financial Economics*, *142*(2), 487–498. https://doi.org/10.1016/j.jfineco.2021.08.004

Tarquinio, L., & Posadas, S. C. (2020). Exploring the term "non-financial information": An academics' view. *Meditari Accountancy Research*, *28*(5), 727–749. https://doi.org/10.1108/MEDAR-11-2019-0602

Tomar, S. (2019). CSR Disclosure and Benchmarking-Learning: Emissions Responses to Mandatory Greenhouse Gas Disclosure. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3448904

UNFCCC. (n.d.). The Paris Agreement. *United Nations Framework Convention on Climate Change*. https://unfccc.int/process-and-meetings/the-paris-agreement

U.S. General Services Administration. (2023). Agency Rule List—Spring 2023. https://www.reginfo.gov/public/do/eAgendaMain?operation=OPERATION_GET_AGENCY_RULE_LIST&currentPub=true&agencyCode=&showStage=active&agencyCd=3235&csrf_token=06A3FFF4C90B5A3674D1B33FCC722CF686BB4CAA705B6B23AF605833884CEA10DE15D9925DC7D86E8DE2B350CDFE2CF1419B

Van der Lugt, C. T., Van de Wijs, P. P., & Petrovics, D. (2020). *Carrots & Sticks 2020—Sustainability reporting policy: Global trends in disclosure as the ESG agenda goes mainstream* (tech. rep.). Global Reporting Initiative (GRI) and the University of Stellenbosch Business School (USB). https://www.carrotsandsticks.net/media/zirbzabv/carrots-and-sticks-2020-june2020.pdf

Verrecchia, R. E. (2001). Essays on disclosure. *Journal of Accounting and Economics*, *32*(1–3), 97–180. https://doi.org/10.1016/S0165-4101(01)00025-8

Wang, L. L. (2023). Transmission Effects of ESG Disclosure Regulations Through Bank Lending Networks. *Journal of Accounting Research*, *61*(3), 935–978. https://doi.org/10.1111/1475-679X.12478

Watts, R. L., & Zimmerman, J. L. (1978). Towards a Positive Theory of the Determination of Accounting Standards. *The Accounting Review*, *53*(1), 112–134.

World Bank. (2023). GDP (current US$)—Germany, France, United Kingdom, European Union. https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?end=2014&locations=DE-FR-GB-EU&most_recent_value_desc=true&start=2012

Junior Management Science 10(3) (2025) 715-747

Junior Management Science

www.jums.academy
ISSN: 2942-1861



Government Interventions During the COVID-19 Pandemic, Culture, and Corporate Cost Behaviour

Meret Anna Gläser

Ludwig Maximilian University of Munich

Abstract

The COVID-19 pandemic triggered unprecedented government interventions, creating a unique setting to examine implications on corporate cost behaviour. This study explores the relationship between the stringency of government interventions during the pandemic and labour cost stickiness, as well as the moderating role of national culture, using 15,446 firm-year observations from 3,383 listed firms across 25 European countries from 2017 to 2022. A difference-in-differences regression analysis reveals that stringent interventions are related to increased labor cost stickiness, suggesting that managers view such measures as a signal of pandemic control which reduces their pessimism about future demand. Additionally, a median-based sample split shows that several dimensions of national culture moderate the relation between governmental stringency and labour cost stickiness, highlighting that culture influences how managers form future expectations based on stringent government interventions. The study connects formal institutions, i.e. governmental interventions, and informal institutions, i.e. national culture, with cost asymmetry as well as expands firm-level cost behaviour research in the context of the COVID-19 crisis.

Keywords: cost stickiness; COVID-19; culture; interventions

1. Introduction

In March 2020 the World Health Organization declared the outbreak of the novel coronavirus disease 2019 (COVID-19) a global pandemic (World Health Organization, 2020). In retrospective, this might have marked the beginning of a crisis which profoundly impacted the lives of millions and disrupted businesses worldwide. To ensure social distancing and curb the spread of the virus, most governments introduced unprecedented interventions, including travel restrictions, stay-at-home orders and school or workplace closures (Hale et al., 2021). Besides the detrimental effects on health, the pandemic, as well as governments' reactions to it, induced severe economic disturbances and confronted companies' decision-makers with extreme levels of uncertainty (Altig et al., 2020; Caggiano et al., 2020). A considerable stream of literature investigates the impact of the outbreak of COVID-19 and subsequent government interventions on aggregate economic metrics like stock returns or exchange rate volatility (e.g. Aggarwal et al., 2021; Feng et al., 2021). However, until today, there is only little research on the microeconomic consequences, such as the relation between government interventions during COVID-19 and corporate cost behaviour.

Traditional cost models assume costs to behave proportionally and symmetrically to changes in activity, independent of the direction of the change (Noreen, 1991). More recent studies suggest an alternative model, arguing that costs are behaving asymmetrically – in other words "sticky" – because managers make the deliberate decision to retain some of the unused resources when activity declines (M. C. Anderson et al., 2003). Next to firm-specific factors, a company's operating environment is found to be associated with resource management decisions and, consequently, cost asymmetry (Banker et al., 2020; Bugeja et al., 2015). Among the various external factors influencing corporate cost behaviour, several studies have highlighted the relationship between

I want to thank Dr. Nina Schwaiger from the Institute for Accounting and Control (IUC) for always providing ideas and impulses to improve my thesis and supporting me throughout the time of writing.

DOI: https://doi.org/10.5282/jums/v10i3pp715-747

© The Author(s) 2025. Published by Junior Management Science.
This is an Open Access article distributed under the terms of the CC-BY-4.0 (Attribution 4.0 International). Open Access funding provided by ZBW.

cost stickiness and both formal and informal country-level institutions (Banker et al., 2013; Holzhacker et al., 2015a; Kitching et al., 2016; Ma et al., 2021).

Given the exogenous nature of the COVID-19-induced changes in activity, along with the varying levels of government interventions' stringency across countries (Finne et al., 2024; Hale et al., 2021), the pandemic provides a unique opportunity to expand research on formal institutions and cost asymmetry. The COVID-19 crisis elicited profound reactions on the European labour market. For instance, the employment rate in the EU decreased by 2 percentage points and the total hours worked fell by 15 percent from the last quarter of 2019 to the second quarter of 2020 (Ando et al., 2022). This highlights why specifically management decisions during COVID-19 regarding labour should be investigated in more detail. Thus, this study leverages the pandemic context to examine the association between the stringency of government interventions during COVID-19 and labour cost stickiness. Although laws and governmental regulations provide the general boundaries for managers' decision-making, informal institutions such as culture impact the decision-making process as well, as they can unconsciously influence managers' values, attitudes, and behaviour (Kitching et al., 2016). Therefore, besides investigating the association between labour cost stickiness and a formal institution, namely government interventions during COVID-19, an additional analysis is conducted to examine the potential role of national culture in moderating this relation. In summary, this study aims at answering the following research questions: *how do government interventions during the COVID-19 pandemic relate to corporate cost behaviour, and does national culture moderate this relation?*

During COVID-19, decision-makers faced unparalleled uncertainty, as the duration of the crisis and its effects on the economy were largely unknown. Based on the findings of previous literature which identify managers' expectations about future sales under demand uncertainty as influencing cost asymmetry (M. C. Anderson et al., 2003; Banker et al., 2014), most authors argue and find that the outbreak of the COVID-19 virus is negatively related to cost stickiness (e.g. BenYoussef et al., 2023; Kwak et al., 2021). The authors claim that the uncertainty surrounding the pandemic increases managers' pessimism about the permanence of sales drops, thus lowering their future demand expectations und consequently also the level of cost stickiness.

Next to the outbreak of the virus itself, governmental reactions to curb its spread might be associated with cost stickiness as well. Government interventions are expected to be related to the level of corporate labour cost stickiness since they can induce managers to reassess their assumptions about the persistence of the crisis and thus change managerial expectations about future demand conditions. However, considering the contradictory argumentation and findings of the literature to date, the direction of this relation is unclear. On the one hand, government interventions might send the signal that the government is able to effectively control the pandemic and therefore reduce uncertainty (e.g. Kizys et al., 2021). This could induce managers to be more optimistic about future demand and consequently to be more inclined to keep slack resources, resulting in increased levels of labour cost stickiness. On the other hand, government interventions might convey a perception of extended economic challenges (BenYoussef et al., 2023) and generate additional uncertainty regarding the interventions' effectiveness and impact (Ashraf, 2020). Thus, corporate decision-makers might be more pessimistic about future demand and therefore be more likely to remove unused resources, which leads to decreased levels of labour cost stickiness.

Moreover, it is expected that national culture leads to one argument predominating the other. National culture influences the understanding and acceptance of formal institutions (Helmke & Levitsky, 2006), which implies that cultural differences across countries may moderate the relation between stringent government interventions during COVID-19 and labour cost stickiness. Specifically, national culture is suggested to affect managers' acceptance of stringent interventions and their belief in the interventions' effectiveness. If managers perceive the stringent interventions as legitimate and effective in mitigating the adverse impacts of the COVID-19 crisis, they may adopt a more optimistic outlook on future demand. This optimism could, in turn, result in higher levels of labour cost stickiness, as managers are more inclined to retain resources in anticipation of a recovery in sales.

This study investigates the research questions using a sample of 15,446 firm-year observations from 3,383 listed firms across 25 European countries between 2017 and 2022. In an initial analysis, the full sample is tested for labour cost stickiness using a model based on the one developed by M. C. Anderson et al. (2003). Additionally, it is examined whether COVID-19 changes the result. Consistent with prior research, the findings indicate that, on average, labour costs exhibit stickiness. However, this stickiness significantly decreases after the outbreak of COVID-19, so that cost asymmetry can no longer be detected in the subsample covering the pandemic years 2020 to 2022. This aligns with the notion that the COVID-19 crisis heightens managerial pessimism, hence keeping corporate decision-makers from retaining unused resources when sales decline. Next, this study exploits the fact that the COVID-19 outbreak, and government interventions following it, can be considered an exogenous shock to firms (Finne et al., 2024). It employs a difference-in-differences (DiD) panel regression model to examine whether there is a relation between the stringency of government interventions during COVID-19 and firms' labour cost stickiness. This design allows comparing the difference in changes in labour cost stickiness from before to after the implementation of stringent government restrictions between countries with high versus low levels of governmental stringency, while controlling for known determinants of cost stickiness. This study finds evidence consistent with the conclusion that stringent government interventions during COVID-19 are associated with significantly higher labour cost stickiness. These results are in line with that stream of the literature, which argues that by introducing stringent interventions, governments

might signal that the pandemic is under control. This, in turn, decreases managerial pessimism and lead managers to retain slack resources during a sales drop. Lastly, to investigate whether an informal institution moderates the relation between a formal institution, namely the stringency of government interventions during COVID-19, and labour cost stickiness, national culture is used as a proxy. Like in other studies which examine cross-country differences in cost stickiness (e.g. Cannon et al., 2020), the sample is split based on the median value of the respective cultural dimension. The baseline model is re-estimated for both groups, and it is determined whether the DiD-coefficient differs between them. The regression estimations indicate that the relation between the stringency of government interventions during COVID-19 and labour cost stickiness varies between high- and low levels of several cultural dimensions, which provides evidence in line with national culture moderating, i.e. amplifying or attenuating, this association.

This study contributes to the existing literature in several ways. It is one of the few studies investigating the association between government interventions during COVID-19 and cost stickiness and it is the first one doing so in the European setting. Moreover, it is the first to examine the potential moderating role of national culture in this relation. Hence, this study adds to the literature on country-level determinants of asymmetric cost behaviour by showing how formal institutions are associated with labour cost stickiness and how informal institutions moderate this relation, in the context of COVID-19. In doing so, it addresses calls to expand research on the impact of governmental actions and cultural factors (Ibrahim et al., 2022) on cost asymmetry. Additionally, this study contributes to the growing research field examining the consequences of the COVID-19 crisis and government interventions related to it. Whereas most studies focus on aggregate, industry-level, or stock return analyses (Finne et al., 2024), this study deepens the research on firm-level consequences of the pandemic.

The remainder of this study is organized as follows. Initially, essential terms and concepts are defined, previous literature is reviewed, and the study's hypotheses are derived. Next, the sample and the research design are described. Subsequently, the results from the regression analyses are presented. In the final section, a discussion of the study's limitations, suggestions for future research, and a conclusion follow.

2. Theoretical perspective and hypotheses development

2.1. Key terms and concepts

2.1.1. Defining asymmetric cost behaviour

Costs are caused by resources that are needed to perform activities within a corporation (Cooper & Kaplan, 1992). In the short run, some resources are hardly adjustable whereas other resources can be flexibly adapted depending on the actual activity level. Accordingly, in the traditional model of corporate cost behaviour costs are classified as either fixed or variable (Brüggen & Zehnder, 2014). It is assumed that fixed costs are constant, while variable costs move proportionally and symmetrically to changes in cost drivers such as production volume or sales (Banker & Byzalov, 2014). The assumption of proportionality and symmetry implies that if variable costs increase for example by 1% for a 1% incline in activity, they will decrease by 1% for an equivalent reduction in activity as well (Calleja et al., 2006). Thus, the average magnitude of the percentage change in cost for each percentage change in activity, i.e. the cost elasticity (Holzhacker et al., 2015a), is equivalent for activity increases and decreases. The traditional model has been extended to consider multiple cost drivers, for example in the activity-based costing (ABC) model[1], and non-linear effects such as economies of scale or learning effects. However, it still postulates a mechanistic relation between costs and cost drivers and does not consider how managerial decision-making may affect the resource adjustment process (Banker et al., 2018).

More recent empirical studies suggest an alternative model. Noreen and Soderstrom (1994, 1997) find initial, however weak, evidence for non-proportional and asymmetric cost behaviour by conducting a cross-sectional and time-series analysis of overhead costs in US hospitals. They conclude that the traditional model overstates the impact of changes in the activity level on concurrent costs. M. C. Anderson et al. (2003) were the first to present statistically and economically significant evidence that corporate cost behaviour is inconsistent with the traditional view of proportionally and symmetrically behaving costs. They show that SG&A costs in a sample of 7,629 US firms during the period 1979 to 1998 increase by 0.55% on average for a 1% sales increase but decrease by just 0.35% on average for a 1% sales decrease. The authors label this phenomenon "cost stickiness" and define costs correspondingly as sticky if "the magnitude of the increase in costs associated with an increase in volume is greater than the magnitude of the decrease in costs associated with an equivalent decrease in volume" (M. C. Anderson et al., 2003, p. 48). They claim that cost stickiness occurs because the resources underlying SG&A costs are not moving mechanistically with changes in sales volume but are actively managed by corporate decision-makers. The authors argue that if managers observe increasing demand, they have to deploy additional resources, as they otherwise would not be able to supply the required volume. However, if observing declining demand, it is at the discretion of the manager to either retain or reduce slack resources. In the case of sticky costs, managers deliberately decide to preserve idle resources because they perceive the retainment as less costly than the downward adjustment (M. C. Anderson et al., 2003). Summarizing, the cause of cost asymmetry is the deliberate management of corporate resources, and concurrent costs, by corporate decision-makers.

[1] In the ABC model costs are allocated to products based on multiple activities which are performed for example on the unit, batch, or product level (Noreen, 1991).

An extensive number of studies build on the seminal paper of M. C. Anderson et al. (2003). Subsequent researchers complement the model of asymmetric cost behaviour by adding the definition of cost anti-stickiness to it. According to Weiss (2010, p. 1142) costs are anti-sticky if "they increase less when activity rises than they decrease when activity falls by an equivalent amount". *Figure 1* summarizes the model of asymmetric cost behaviour and highlights the differences between symmetric, sticky, and anti-sticky costs.

If symmetric costs are assumed, the slope of the cost function (C) is linear and therefore identical for high (Y_H) and low activity levels (Y_L), for example sales increases and decreases. In the case of sticky costs, the slope is flatter for low activity levels (Y_L) than for high activity levels (Y_H). When the cost function exhibits anti-sticky behaviour, the slope is steeper for low activity levels (Y_L) than for high activity levels (Y_H). This implies that in the case of sticky (anti-sticky) costs, the cost elasticity is greater (lower) for sales increases than for sales decreases.

Subsequent research uses the regression model of M. C. Anderson et al. (2003) as a basis to provide evidence of asymmetric cost behaviour across different cost categories[2], countries[3] and time periods. Whilst a comparably small number of authors investigates the consequences of asymmetric costs (for a summary see Ibrahim et al., 2022), a comprehensive stream of literature examines the determinants of cost asymmetry.

2.1.2. Characterizing determinants of asymmetric cost behaviour

The firm- and country-level determinants of asymmetric cost behaviour identified by prior literature can be classified into three main categories (Banker et al., 2018), namely, managers' expectations for future sales, resource adjustment costs and managerial opportunistic motives. They represent the channels through which cost asymmetry can be affected. As will be detailed in the following sections, it is assumed that the stringency of government interventions during COVID-19 is associated to cost asymmetry through the channel of managerial future expectations. Additionally, national culture is expected to moderate this relationship via the same channel. Therefore, the primary focus is on the channel of managerial future expectations. However, the literature on the channels of adjustment costs and opportunistic managerial motives is acknowledged for completeness.

Managerial decision-making and consequently cost behaviour is influenced not only by concurrent changes in demand but also by managers' expectations about future sales levels (Banker et al., 2018, 2020). Resource management decisions are made under conditions of demand uncertainty (Banker et al., 2014), requiring managers to evaluate the permanence of a sales decline when deciding whether to retain or cut unused resources. Some authors argue that demand uncertainty is associated with increased cost stickiness. They suggest that managers purposely delay the resource adjustment decision and keep slack resources until they can gather more information about the permanence of the sales decline (e.g. M. C. Anderson et al., 2003; W. J. Lee et al., 2020). This materializes in an increased level of cost stickiness which reverses in subsequent periods when more information on the duration of the sales drop is known and uncertainty resolves. Other authors expand this argument and claim that the effect also depends on managerial *expectations* about the persistence of the demand uncertainty. If managers are optimistic and believe that the sales drop is temporary, they are more likely to keep unutilized resources, as the expected retention costs are smaller than the expected adjustment costs, resulting in greater cost stickiness. Conversely, if managers are pessimistic and believe the sales drop is permanent, the retention costs are expected to be higher than the adjustment costs, leading to the cutting of unused resources and, consequently, to less cost stickiness (Banker & Byzalov, 2014; Kama & Weiss, 2013). This reasoning applies to sales increases as well: optimistic managers are more willing to expand resources as they are more likely to classify the increase as permanent, while pessimistic managers are hesitant to commit resources due to the fear of having to reverse these commitments (Banker & Byzalov, 2014). The influence of managers' expectations on cost stickiness is first mentioned by M. C. Anderson et al. (2003). They show that the re-occurrence of declining sales in two consecutive periods is associated with increased pessimism about the permanence of the demand reduction and consequently decreasing SG&A cost stickiness, whereas macroeconomic growth is related to increased optimism about the duration of the sales decline and consequently increasing SG&A cost stickiness. Further empirical evidence is provided by Banker et al. (2014), who demonstrate that managers tend to extrapolate past trends, leading to more optimistic (pessimistic) expectations about future sales following a prior sales increase (decrease) and resulting in increased (decreased) cost stickiness accordingly. Their findings are reinforced and expanded by Ciftci and Zoubi (2019) who find that in addition to the sales change's direction, also the magnitude of the change influences managers' expectations and thus cost asymmetry.[4] Next to rational, economic information about future sales, managers' optimism can also be based on psychological and behavioural biases like overconfidence (Banker & Byzalov, 2014). Literature suggests that overconfidence fosters excessive optimism about future sales, resulting in increased cost stickiness. As proposed, several studies find a positive rela-

[2] In addition to SG&A costs, operating costs (e.g. Kama and Weiss, 2013), costs of goods sold (e.g. Ibrahim, 2015), labour costs (e.g. Prabowo et al., 2018) and total costs (e.g. Subramaniam and Weidenmier Watson, 2016) are found to exhibit asymmetric cost behaviour.

[3] Prior literature finds evidence for asymmetric cost behaviour in cross-country samples (e.g. Banker and Byzalov, 2014; Calleja et al., 2006), including developed as well as developing countries.

[4] Their results show that cost stickiness is greater (smaller) for small (large) decreases in sales, as managers assess the permanence of the decline as temporary (permanent) and therefore are more optimistic (pessimistic) about future sales changes (Ciftci & Zoubi, 2019).

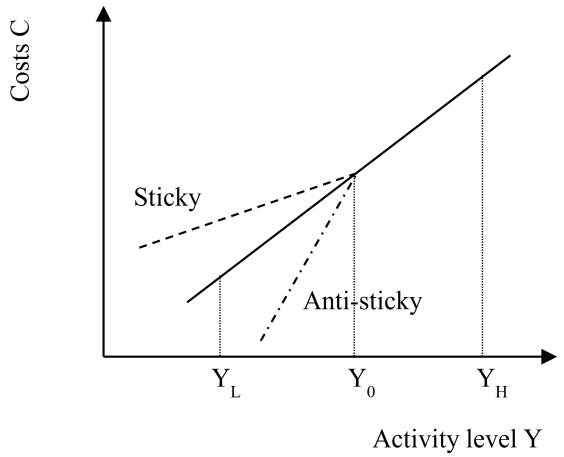


Figure 1: Graphical representation of the cost asymmetry model (Own figure based on Weiss (2010, p. 1444).)

tion between proxies for managerial overconfidence and cost stickiness[5].

Moreover, cost management decisions are influenced by resource adjustment costs, which are defined as costs "to remove committed resources and to replace those resources if demand is restored" (M. C. Anderson et al., 2003, p. 49). Adjustment costs include for example severance payments when cutting resources due to falling activity, or hiring costs, when replacing resources due to rising activity. However, they are not only of monetary nature, but also encompass implicit costs such as a loss of morale or efficiency if teamwork is disrupted (Banker & Byzalov, 2014). The higher the expected resource adjustment costs, the more reluctant managers are to cut excess resources when sales decline, and, in turn, the higher the level of cost stickiness. Several studies examine the relation between proxies for adjustment costs and cost stickiness. For example, Balakrishnan and Gruca (2008) provide department-level findings in showing that cost stickiness is higher for core functions than for ancillary and support services, since the adjustment costs are higher for core activities. M. C. Anderson et al. (2003) suggest that firms with higher asset or employee intensity incur higher adjustment costs as they lose more firm-specific investment when dismantling customized machinery or incur higher severance payments when dismissing employees, respectively. Accordingly, they find a positive association between asset or employee intensity and cost stickiness. Other studies show that country-level proxies for adjustment costs, such as the strictness of employment protection legislation, are related to higher levels of labour and SG&A cost stickiness (Banker et al., 2013).

Lastly, resource adjustment decisions are affected by managerial opportunistic motives. In contrast to studies from the previously described categories, which assume that managers act in the firm's best interest (Brüggen & Zehnder, 2014), this body of literature builds on the agency theory[6] and conjectures that managers make self-interested decisions at the expense of firm value. As self-interested managers want to maximize their private benefit, they may engage in opportunistic behaviours like empire-building. These managers try to maximize the company's size and consequently are more reluctant to cut slack resources, even if sales decline. Hence, the degree of cost stickiness is increased beyond its optimal level resulting from economic factors, i.e. adjustment costs and future expectations (C. X. Chen et al., 2012), resulting in deteriorating firm value (Banker et al., 2018). Authors substantiated this theory by showing that corporate governance diminishes the positive relation between empire-building incentives and cost stickiness as it reduces the agency problem through effective monitoring (Bugeja et al., 2015; C. X. Chen et al., 2012). Next to managers' intrinsic motives, corporations can incentivize managers with performance compensation (Banker et al., 2018). For example, managers' incentives to avoid losses or reach earnings targets (Dierynck et al., 2012; Kama & Weiss, 2013) are also found to be related to resource management decisions and consequently cost asymmetry.

2.1.3. Defining institutions

Managerial decisions "do not emerge independently of the social environment but are jointly shaped by formal and informal institutions" (Hartlieb et al., 2020b, p. 26). Thus, researchers acknowledge that cross-country differences in cost stickiness might exist due to regulatory or cultural aspects (Stimolo & Porporato, 2019) and express interest in

[5] For example, C. X. Chen et al. (2022) find a positive association between CFO overconfidence, proxied by option-exercising behaviour, and cost stickiness. Additionally, D. Yang (2015) shows that CEO hubris is positively related to cost stickiness in examining overconfident bidder CEOs in merger events.

[6] The agency theory assumes that, in a situation of separation of ownership and control, the principal (i.e. a corporation's shareholders) and the agent (i.e. corporate managers) are both utility maximisers. Thus, an "agent will not always act in the best interests of the principal" (Jensen & Meckling, 1976, p. 308).

understanding the relation between cost management decisions and formal and informal country-level institutions (e.g. Banker et al., 2013; Kitching et al., 2016).

Institutions are defined as human-made "systems of established and embedded social rules that structure interactions" (Hodgson, 2006, p. 18) by "constraining and enabling actors' behaviour" (Helmke & Levitsky, 2006, p. 5). They allow individuals to form stable expectations about the behaviour of others and thus make ordered human interaction possible (Hodgson, 2006). Both formal and informal institutions can be summarized under this definition (Redmond, 2005), yet they differ in how they are transmitted and enforced.

2.1.4. Characterizing government interventions during COVID-19 as formal institution

Redmond (2005, p. 666) defines formal institutions as "administered by a central authority" that has the power "to set rules which will direct the behaviour of many, as well as [...] to interpret and enforce the rules". These explicit and codified rules and procedures are "communicated and enforced through channels that are widely accepted as official" (Helmke & Levitsky, 2006). Formal institutions comprise for example political and economic rules like laws and contracts (North, 1990) which set the boundaries for managerial decision-making, including resource and cost management decisions (Kitching et al., 2016).

Prior literature shows that cost behaviour varies with country-level formal institutions. For instance, Banker et al. (2013) investigate the impact of legal institutions on cost stickiness and report that cost stickiness is increasing in the strictness of employment protection legislation as the costs of firing employees, i.e. the downward adjustment costs, rise. Holzhacker et al. (2015a) and Belina et al. (2019) examine how changes in governmental healthcare regulations affect cost asymmetry and find that cost stickiness in German hospitals and US health insurers decreases due to rising operational risk and cost downward pressure introduced by the regulatory adjustments. The results of both Jin and Wu (2021) and W. J. Lee et al. (2020) reveal that political institutions moderate the relationship between political uncertainty and cost stickiness. Whilst the former show that high government spending, low property rights and low levels of labour regulation strengthen the proposed negative association between political uncertainty and cost stickiness, the latter report evidence that stable political and legal institutions reinforce cost stickiness in election years. Kuo and Lee (2023) show that strong political institutions are related to increased levels of cost stickiness, as they strengthen contract enforcement, corruption control and judicial independence.

This study utilizes government interventions implemented during the COVID-19 pandemic as a proxy for formal institutions, as they serve as a contemporary and globally relevant example of how formal institutions can exert influence across different countries. Governments during the pandemic acted as a central authority, setting up and enforcing codified rules to guide and constrain human behaviour. To ensure social distancing and curb the spread of the virus, most governments introduced interventions, including for example school or workplace closings, travel restrictions or stay-at-home requirements. Though, the degree of individual governments' interventions in Europe differed substantially, ranging from advisory measures to mandatory rules (Hale et al., 2021). For instance, some countries, including Italy and Germany, implemented mandatory stay-at-home requirements, while others, such as Sweden, only recommended working from home (Akter, 2020). This variance of stringency makes governmental interventions during COVID-19 a beneficial case to examine the relation between formal institutions and corporate cost stickiness. Although governments also issued health system policies, like testing or vaccination guidelines, and offered economic aid such as income support or debt reliefs (Hale et al., 2021), this study follows BenYoussef et al. (2023) and only considers non-pharmaceutical, non-financial government interventions in order to provide more nuanced insights.

2.1.5. Characterizing culture as informal institution

Informal institutions are defined as "socially shared rules, usually unwritten, that are created, communicated and enforced outside officially sanctioned channels" (Helmke & Levitsky, 2006, p. 3) and include for example "norms, conventions, customs and traditions" (Redmond, 2005, p. 666). They define acceptable or desirable behaviour which individuals comply to, as they try to avoid costs of deviant behaviour such as the feeling of embarrassment or shame (Sunstein, 1996).

Several studies show that various types of informal institutions are related to cost asymmetry. For instance, Ma et al. (2021) find that religion is negatively associated with corporate cost stickiness, as religious conservatism and ethicality tend to reduce managers' (over-)optimism and empire-building tendencies. Hartlieb et al. (2020a) state that community social capital, i.e. cooperative norms and social networks, limits managers' pursuit of opportunistic motives and therefore mitigates sticky cost behaviour. In another study, Hartlieb et al. (2020b) show that generalized trust has a positive effect on cost stickiness by increasing psychological adjustment costs and managerial optimism. Kitching et al. (2016) find that several dimensions of national culture, namely uncertainty avoidance, masculinity and long-term orientation, are related to operating cost stickiness through the managerial expectation channel and the adjustment cost channel. They argue that cost stickiness is less pronounced in firms located in countries being high in uncertainty avoidance and long-term orientation, as managers of such firms are more loss-averse and thrifty, respectively. Moreover, cost stickiness is attenuated in firms located in countries high in masculinity due to managers' lower psychological downward adjustment costs.

The aforementioned definition of informal institutions also includes, but is not restricted to, national culture (Helmke & Levitsky, 2006; North, 1990). Given that managers frequently operate at corporate headquarters (Ma et

al., 2021), it is assumed that they are, to some extent, part of the national society of the country where the firm's headquarters is located. National culture is therefore expected to impact managerial resource decision-making by forming managers' values, beliefs, and assumptions (Kitching et al., 2016). Thus, this study uses national culture as defined by Hofstede et al. (2010) as a proxy for informal institutions. Hofstede et al. (2010, p. 6) refer to culture as "the collective programming of the mind that distinguishes the members of one group [...] from others". The authors explain that culture is learned through socialization processes and state that specifically nations are a "source of a considerable amount of common mental programming" (Hofstede et al., 2010, p. 21). They propose that national culture varies along six dimensions, namely power distance (PDI), individualism (IDV), masculinity (MAS), uncertainty avoidance (UAI), long-term orientation (LTO) and indulgence (IVR) and provide definitions for each of them. *Power distance* is defined as the degree to which unequal distribution of power, status and authority is accepted by lower status members of the society. In countries high in *individualism*, decisions are made based on individual needs, whereas in collectivistic countries, people are integrated into strong, cohesive groups and thus put a larger emphasis on protecting group interests. In societies characterized by high *masculinity*, social gender roles are clearly separated, meaning that "men are supposed to be assertive, tough, and focused on material success, whereas women are supposed to be more modest, tender, and concerned with the quality of life" (Hofstede et al., 2010, p. 140). Contrastingly, in feminine societies, social gender roles are suggested to overlap. *Uncertainty avoidance* is defined as "the extent to which the members of a culture feel threatened by ambiguous or unknown situations" (Hofstede et al., 2010, p. 191). While societies with a *long-term orientation* foster virtues oriented towards future rewards, like perseverance and thrift, short-term oriented societies support virtues related to the past and present, such as tradition and fulfilling social obligations. Lastly, *indulgence* is defined as "a tendency to allow relatively free gratification of basic and natural human desires related to enjoying life and having fun", whereas restraint is the "conviction that such gratification needs to be curbed and regulated by strict social norms" (Hofstede et al., 2010, p. 281).

2.2. Review of prior literature and hypothesis development

2.2.1. COVID-19 and asymmetric cost behaviour

On average, labour costs are expected to be sticky. This is confirmed for various samples in prior studies (e.g. Dalla Via and Perego, 2014; Dierynck et al., 2012; Prabowo et al., 2018). However, as Banker et al. (2018, p. 192) state: "costs that are sticky on average are not always sticky". Corporate cost stickiness indeed depends on various factors. In addition to firm-specific characteristic, the external environment, for example in terms of macroeconomic conditions, affects managers' operating decisions and hence the degree of cost stickiness (Banker et al., 2020; Bugeja et al., 2015). This study focuses on examining the relationship between cost stickiness and formal institutions, i.e. government interventions, and the potential moderating role of informal institutions, i.e. national culture, in a specific context, namely the COVID-19 crisis. The pandemic caused unexpected macroeconomic disruptions, including severe demand drops[7], which are likely to be related to resource management decisions. Hence, to delineate the setting and to establish the foundation for subsequent argumentations and analyses, the current state of research on cost stickiness and COVID-19 is reviewed and considered.

Prior research suggests that cost stickiness is decreasing during times of crisis. Banker et al. (2020) provides several reasons why the economic slowdown induced by a crisis is related to lower levels of cost stickiness. Firstly, as external financing becomes more expensive, managers may want to reduce costs to preserve cash. Furthermore, managers might have more pessimistic sales expectations, leading them to avoid preserving slack resources. Lastly, under the existential threat of dropping out of business, managers might be less resistant to operational changes which could improve organizational efficiency, thus being more inclined to adjust redundant resources downward. Additionally, literature suggests that managers are more likely to avoid committing contracts during periods of crisis to increase flexibility. For example, managers prefer hiring temporary workers which can be dismissed relatively easily (Ibrahim, 2015). Due to the contractual arrangement, firing costs for temporary employees are lower, resulting in less resource adjustment costs (Banker et al., 2013) and consequently decreasing cost stickiness. Economic downturns lower future adjustment costs as well, as it is easier to re-negotiate input costs such as wage rates (Jin & Wu, 2021).

The financial crisis of 2008 is found to induce cost anti-stickiness in samples of Egyptian firms (Ibrahim, 2015), UK Chemical firms (Hassanein & Younis, 2020) and municipalities in Spain (Karatzimas et al., 2022). Similarly, prior research shows that the COVID-19 pandemic is related to a decrease cost stickiness. COVID-19 confronted companies with unparalleled economic uncertainty, especially as the duration of the crisis was largely unpredictable in the absence of a similar historical event (Altig et al., 2020). As mentioned before, the level of cost stickiness depends on managers' expectations about the duration of the prevailing uncertainty and their optimism or pessimism about the permanence of a sales drop. In line with this argument, most studies in the field propose and find a negative association between COVID-19 and cost stickiness. The authors claim that the pandemic-induced un-

[7] For completeness, it should be acknowledged that COVID-19 significantly affected both the demand and the supply side of businesses. Interpreting COVID-19 as a supply instead of a demand shock likely has different implications for cost stickiness. Scarcity of input resources may lead to higher adjustment costs, prompting managers to retain unused resources and resulting in increased cost stickiness. However, since firms generally perceive the pandemic primarily as a demand shock (Meyer et al., 2022), and this perspective is supported by prior literature, this study adopts the same viewpoint.

certainty caused managers to form pessimistic expectations about consumers' purchasing power and the prospect of sales rebounding in the short-term (Meyer et al., 2022). Therefore, managers are likely to assess the retention of resources to be more costly than the resource downward adjustment. Consistently, results show lower levels of cost stickiness during the pandemic. For example, Buchheim et al. (2022), although not directly examining cost stickiness, detect that firms who expect the duration of the shutdown to be longer than four months, cut significantly more resources, for example in terms of layoffs or the cancellation of investments. A negative association of the COVID-19 outbreak with cost stickiness is furthermore found in smaller samples of IT companies (Kwak et al., 2021) and healthcare firms (Ekici et al., 2024), as well as larger samples of American publicly listed companies (BenYoussef et al., 2023; Ghazy et al., 2024).

2.2.2. Government interventions during COVID-19 and asymmetric cost behaviour

When examining the consequences of COVID-19, it is important to differentiate between the effect of the pandemic itself and the government regulations implemented in response to it. This also applies when researching cost stickiness, as the study of Y. Yang and Chen (2024) shows. The authors investigate the influence of COVID-19 on cost stickiness in China and Australia and find the relation between the pandemic and cost stickiness to differ between both countries. They suggest that dissimilarities in policy responses to COVID-19 might explain this variation. Thus, to facilitate a clear distinction, this study focuses on the direct relation between the stringency of government interventions during COVID-19 and labour cost stickiness.

Literature examining government interventions during COVID-19 in the field of cost stickiness is scarce. The working papers by BenYoussef et al. (2023) and Ghazy et al. (2024) constitute notable exceptions. BenYoussef et al. (2023) examine SG&A cost behaviour in a North American sample and find a negative association between governmental stringency during COVID-19 and cost stickiness. Ghazy et al. (2024) investigate the relation between the strictness of workplace closures throughout the pandemic, as well as governmental economic support, and operating cost stickiness in US firms. Whereas their results show that cost stickiness exhibits a negative association with the strictness of workplace closure, economic support does not have a statistically significant effect. Due to the lack of theory regarding government interventions during COVID-19 in cost stickiness research, this study relies on argumentations and findings from another field, namely the finance literature. As Ashraf (2020, p. 1) states, stock markets "provide an incentivized survey of future expected outcomes". This makes the literature examining the impact of governmental stringency during COVID-19 on stock markets a suitable source for developing this study's theoretical framework, assuming that investors' and managers' future expectations are comparable.

Generally, the stringency of government interventions during COVID-19 is expected to be associated to labour cost stickiness by changing managers' perception of the COVID-19-induced uncertainty and relatedly their expectations about the permanence of the economic downturn. In other words, it is suggested to be related to cost stickiness through the channel of managerial expectations about future demand, which was introduced in section 2.1.2. However, considering the results of the literature to date, the direction of the association between the stringency of government interventions during COVID-19 and cost stickiness remains unclear.

On the one hand, government interventions might send the signal that the pandemic is under control and therefore reduce uncertainty (Caggiano et al., 2020; Kizys et al., 2021). Several studies present evidence in line with this suggestion by showing a positive relation between government interventions and stock market return (Saif-Alyousfi, 2022; H. Yang & Deng, 2021). In addition to the research on stock markets, Feng et al. (2021) show that government non-pharmaceutical interventions are associated with decreasing exchange rate volatility as they send positive signals to markets and investors, thus effectively reducing panic and uncertainty. Like investors and markets in general, also corporate managers may interpret stringent government interventions as effectively reducing COVID-19-induced uncertainty. This might result in managers being more optimistic about demand recovery and consequently more inclined to keep slack resources, resulting in increased cost stickiness.

On the other hand, government interventions may give rise to additional uncertainty, for example regarding the interventions' duration and effectiveness (Altig et al., 2020; Ashraf, 2020). For instance, Zaremba et al. (2020) identify a positive association between the stringency of policy responses and stock market volatility. By accounting for country-specific factors and pandemic affectedness, their study demonstrates that the stringency of government interventions is an independent driver of increased volatility, distinct from the effects of the pandemic itself. Moreover, stringent government interventions might convey a perception of extended economic challenges (BenYoussef et al., 2023). Supporting this view, Ashraf (2020) finds that the announcement of stringent interventions is negatively related to stock market returns, as it leads to more pessimistic expectations due to the interventions' expected adverse impact on economic activity. In summary, stringent government interventions might as well lead corporate decision-makers to be more pessimistic about future demand. Hence, managers might be more likely to remove resources, which results in decreasing cost stickiness, as found by BenYoussef et al. (2023) and Ghazy et al. (2024).

Consistent with this double-sided argumentation, Aggarwal et al. (2021) show in the context of stock markets that the impact of government interventions' stringency during COVID-19 on investor behaviour is twofold. Their results suggest that more stringent government interventions positively influence the overall returns by making investors feeling more secure, as reflected in the declining equity risk premia demanded. On the other hand, the results also depict

that increased stringency of government interventions is negatively related to the overall return as it lowers investors' growth estimates. Since theory and prior literature do not clearly predict the direction of the association between stringent government interventions during COVID-19 and labour cost stickiness, it remains an open empirical question. Furthermore, it could also be the case that the proposed opposite effects of stringent government interventions might cancel each other out. Thus, the hypothesis is stated in a neutral form:

> *Hypothesis 1: There is no relation between the stringency of government interventions during COVID-19 and cost stickiness.*

2.2.3. National culture as moderator

As prior research emphasizes, both formal and informal institutions influence decision-making, which highlights the importance of studying them in combination. For example, Graafland and Noorderhaven (2020, p. 1040) state that "[...] studies should not only look at both institutions and culture, but also at the interactions between these types of factors." Nevertheless, until today there are only few studies in the field of cost asymmetry research following this call. One exception is Cannon et al. (2020), who do not only examine the direct effect of takeover laws on SG&A cost stickiness, but also test whether culture, in terms of short-termism, moderates this relation. Their findings reveal that the negative relation between the passage of takeover laws and cost stickiness is more emphasized in short-term oriented countries.

Similarly, it is expected that national culture moderates the relation between the stringency of government interventions during COVID-19 and labour cost stickiness. National culture influences the understanding and acceptance of formal institutions (Helmke & Levitsky, 2006), which is why culture is also suggested to impact societal beliefs about the legitimacy of governments' mitigation measures (Wang et al., 2022). This view is supported by Dheer et al. (2021, p. 1873), who state that culture shapes "how individuals interpret and comply with stringent government measures". Generally, this might apply to all of the six cultural dimensions defined by Hofstede et al. (2010). For brevity, the reasoning is exemplary shown for only two of them.

For instance, individuals in collectivistic cultures might be more likely to acknowledge the legitimacy of stringent government interventions during COVID-19, as they are more willing to sacrifice individual freedom for the common good (Biddlestone et al., 2020; Dheer et al., 2021). This is supported by previous literature claiming that the state has a dominant role in collectivistic nations (Hofstede et al., 2010) and that individuals in collectivistic cultures derive satisfaction from fulfilling social obligations (Inglehart & Oyserman, 2004). In line with this, J. Lee et al. (2024) find that collectivistic cultures are more likely to endorse stringent government policies during COVID-19, whereas individualistic cultures are more prone to reject them.

In contrast, nations high in uncertainty avoidance might be less likely to accept stringent government interventions during COVID-19, as such measures are perceived as a source of additional uncertainty, for example regarding their economic impact (Altig et al., 2020; Ashraf, 2020). In addition, stringent interventions require people to profoundly change their everyday private and business life, due to measures like stay-at-home requirements or workplace closures. This can be particularly unsettling in uncertainty-avoidant cultures, where change is perceived as evoking unpredictability. Bakry et al. (2022) confirms this argument by showing that in emerging markets, which tend to have higher uncertainty avoidance scores, there is a positive association between the stringency of government interventions and stock volatility. Consistently, individuals from countries high in uncertainty avoidance are found to be less willing to comply to stringent interventions during COVID-19 (Dheer et al., 2021).

To summarize, the stringency of government interventions during COVID-19 is claimed to be related to labour cost stickiness in one of two opposing ways, as elaborated on in the previous section 2.2.2. Stringent measures might either increase managers' optimism about future demand and thus lead to rising levels of cost stickiness or lower managerial expectations about future demand, resulting in decreasing levels of cost stickiness. National culture is expected to be the factor determining which of both forces overweighs. In particular, culture is suggested to influence whether a national society, and thus also managers socialized within it, believes in the legitimacy of stringent government interventions during COVID-19. The acceptance or rejection of stringency is, in turn, believed to shape managers' future demand expectations. If managers accept stringent government interventions, they might believe in their effectiveness in reducing part of the COVID-19-induced uncertainty, which leaves them more optimistic about the temporary nature of declining sales. Ultimately, this materializes in higher levels of cost stickiness. Though, if managers reject stringent government interventions, they might doubt their usefulness in mitigating the adverse effects of the pandemic. They consequently are more pessimistic about future demand conditions, which results in lower cost stickiness. Thus, the extent to which managers' expectations about future demand are shifted due to government interventions during COVID-19 is likely to differ contingent on cultural factors (Buchheim et al., 2022).

Managers' likelihood to accept (reject) stringent government interventions, their resulting optimism (pessimism) about future demand and consequently the extent of cost stickiness exhibited, depend on the specific cultural dimension considered. Hence, the direction of the moderating role of culture in the relation between the stringency of government interventions during COVID-19 and cost stickiness differs between the various cultural dimensions. Therefore, the hypothesis is stated in a neutral form:

> *Hypothesis 2: National culture moderates the relation between government interventions during COVID-19 and cost stickiness.*

3. Data and methodology

3.1. Sample construction

This study is based on panel data from the Thomson Reuters Eikon database, including observations of European publicly listed firms from 2017 to 2022[8]. The firm-level data is merged with country-level data on gross domestic product (GDP) growth obtained from the World Bank website[9], data on the stringency of governmental interventions during COVID-19 from the Oxford Government Response Tracker website[10] as well as data on national culture from Geert Hofstede's website[11]. The sample construction procedure largely follows prior cost stickiness research and is summarized in *Table 1*. Only non-duplicate firm-year observations (S. W. Anderson & Lanen, 2007) from non-financial and non-utility firms[12] are considered (Hartlieb et al., 2020a). Observations with missing or negative labour costs or sales revenue in the current or two preceding years (Banker et al., 2013), as well as observations with missing values for control variables (Hartlieb et al., 2020a) and observations for which the ln-specification cannot be calculated, are deleted. In contrast to most prior studies (e.g. M. C. Anderson et al., 2003), observations in this study are not dropped if costs exceed sales revenue. The observations are only discarded if labour costs exceed sales revenue in the current *and* the prior year, following a suggestion of S. W. Anderson and Lanen (2007).[13] Finally, observations are removed if there are fewer than 50 firm-year observations per country, to ensure that the model is reasonable for each country individually (Hartlieb et al., 2020b). To make the variables comparable between countries with differing currencies, the financial data is converted to euro[14]. All continuous variables are winsorized at the 1st and 99th percentile to weaken the impact of extreme outliers. The final sample contains 15,446 firm-year observations from 3,383 firms in 25 European countries[15].

The sample distribution is presented in Panels A, B and C of *Appendix 1*. Although the number of observations per year rises, the distribution around the COVID-19 outbreak year 2020 is roughly symmetrical, with 43.9% of the observations being in the pre-COVID-19 period and 56.1% in the COVID-19 period. Most of the firms in the sample are headquartered in the United Kingdom (20%)[16], followed by Germany (14%) and France (12%). This is consistent with these three countries being the largest in Europe resident-wise and speaks for the representativeness of the sample. Most of the firms belong to the manufacturing (44%) and the services industry (28%).

3.2. Variable measures

Stringency of government interventions during COVID-19

The stringency of government interventions during COVID-19 is captured by the stringency index (SI). The SI measures the level of non-pharmaceutical, non-financial actions that a government takes to contain the spread of the corona virus, mainly by reducing individuals' mobility.[17] It is obtained from the publicly available Oxford COVID-19 Government Response Tracker (OxCGRT) database from the Oxford University's Blavatnik School of Government, which contains data on government responses during the pandemic for more than 150 countries from January 2020 to December 2022. The SI is based on the unweighted addition of nine closure and containment indicators[18] and is scaled to vary from 0 to 100, with higher values indicating stricter government responses (Hale et al., 2021). It is assumed that only the stringency of government interventions in the country of a firm's headquarters is relevant for managers' decision-making, even if the company operates in other countries as well. As mentioned before, the SI varies significantly across European countries, which is illustrated by *Appendix 2*.

National culture

As the dimensions of national culture from Hofstede et al. (2010) are widely accepted and applied in empirics (Dheer et al., 2021; Kitching et al., 2016), the six dimensions power distance, individualism, masculinity, uncertainty avoidance, long-term orientation and indulgence are used in this study to measure national culture. Each country is assigned a score

[8] Constructing the control variable $SUC_DEC_{i,t}$ requires data from two additional years. However, the years 2015 and 2016 are excluded during the sample selection process, resulting in a symmetrical time span from 2017 to 2022, centred around the outbreak of COVID-19 in the beginning of 2020.

[9] https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG.

[10] https://github.com/OxCGRT/covid-policy-tracker.

[11] https://geert-hofstede.com/national-culture.html.

[12] Firms in the financial and utility industries are dropped to ensure comparability. The Standard Industrial Classification Code (SIC-Code) is used to assign firms to an industry.

[13] Banker et al. (2014) suggest excluding observations where costs exceed sales, as these may reflect unusually high resource commitments. However, the COVID-19 pandemic caused significant declines in sales (Finne et al., 2024), which could result in labour costs exceeding sales as well – not due to excessive costs, but because of decreasing sales revenue. Consequently, removing these observations might bias the analysis. The empirical results remain similar, if observations with labour costs exceeding sales are not discarded at all, as suggested by Banker and Byzalov (2014).

[14] The empirical results remain robust when adjusting for inflation by deflating the financial variables with country-specific GDP deflators.

[15] Since data is not available for all firms over the entire sample period, the longitudinal panel is unbalanced. However, this is consistent with prior literature. For instance, the sample of M. C. Anderson et al. (2003) contains approximately 8.5 observations per firm in a sample period covering 20 years.

[16] According to Finne et al. (2024), it is common in the United Kingdom that the fiscal year ends in March. To examine whether this biases the analysis, a robustness check is conducted. Assigning all firm-year observations with a fiscal year end date before the 30th of June to the previous year does not change the results.

[17] However, the stringency index contains no information on the effective implementation or the appropriateness of the interventions (Hale et al., 2021).

[18] Namely school closings, workplace closings, cancellation of public events, restrictions on gathering size, closing public transport, stay-at-home requirements, restrictions on national movement, restrictions on international travel and public information campaigns (Hale et al., 2021).

Table 1: Sample selection

	Number of observations
Initial sample (2015-2022)	82,504
Less:	
(1) Observations with duplicate ISINs	8
(2) Observations of financial firms (2-digit SIC codes 60-69), public utilities (2-digit SIC code 49) or firms with non-specified industry (no SIC code)	37,536
(3) Observations with missing labour costs (sales revenue) in the current or (two) preceding year(s)	21,431
(4) Observations with negative labour costs (sales revenue) in the current or (two) preceding year(s)	42
(5) Observations where current labour costs exceed sales revenue of the current and the preceding year	2,420
(6) Observations with missing values for other required variables	5,445
(7) Observations from countries with less than 50 firm-year observations	176
Final sample (2017-2022)	15,446

between 0 and 100 for each dimension, with a higher score implying a stronger cultural parameter. Following Kitching et al. (2016), it is assumed that the national culture of the country in which the firm is headquartered is representative for the cultural values of its corporate decision-makers.

3.3. Model specification

Basic model

In a first step, an equation based on the ln-linear[19] model established by M. C. Anderson et al. (2003) is employed to test for the presence of labour cost stickiness. Whereas M. C. Anderson et al. (2003) examine SG&A costs, this study focuses on labour costs, as they represent a significant proportion of total costs and are susceptible to managerial discretion (Pinnuck & Lillis, 2007). Furthermore, examining one specific type of cost potentially allows to locate in which cost category cost stickiness originates (Ibrahim et al., 2022). Following prior literature on labour cost stickiness, changes in the activity level are proxied by the *Net Sales or Revenues* item from Datastream, whereas changes in the labour cost level are operationalized using the *Salaries and Benefits Expenses* item from Datastream (Prabowo et al., 2018). Like Banker et al. (2013), Hartlieb et al. (2020a) and Ma et al. (2021), this study utilizes an extended version of the M. C. Anderson et al. (2003) model. Thus, the equation includes not only the three-way interaction terms of the control variables with $\Delta \ln SALES_{i,t} x DEC_{i,t}$, but also their two-way interactions with $\Delta \ln SALES_{i,t}$, as the control variables are found to be related to cost elasticity independent of a decrease in sales as well (Holzhacker et al., 2015a)[20]. Model (1) shows the corresponding empirical regression equation:

$$\begin{aligned}\Delta \ln LABOUR_{i,t} = {} & ß_0 + ß_1 \Delta \ln SALES_{i,t} \\ & + ß_2 DEC_{i,t} x \Delta \ln SALES_{i,t} \\ & + ß_3 \Delta \ln SALES_{i,t} x SUC_DEC_{i,t} \\ & + ß_4 \Delta \ln SALES_{i,t} x DEC_{i,t} x SUC_DEC_{i,t} \\ & + ß_5 \Delta \ln SALES_{i,t} x GDP_{n,t} \\ & + ß_6 \Delta \ln SALES_{i,t} x DEC_{i,t} x GDP_{n,t} \\ & + ß_7 \Delta \ln SALES_{i,t} x AINT_{i,t} \\ & + ß_8 \Delta \ln SALES_{i,t} x DEC_{i,t} x AINT_{i,t} \\ & + ß_9 \Delta \ln SALES_{i,t} x EINT_{i,t} \\ & + ß_{10} \Delta \ln SALES_{i,t} x DEC_{i,t} x EINT_{i,t} \\ & + \varepsilon_{i,t}\end{aligned} \tag{1}$$

where $\Delta \ln LABOUR_{i,t}$ represents the ln-change in labour costs for firm i from year t-1 to year t and $\Delta \ln SALES_{i,t}$ displays the ln-change in net sales for firm i from year t-1 to year t. $DEC_{i,t}$ is an indicator variable which equals one if sales of firm i decrease from year t-1 to year t. Thus, the coefficient $ß_1$ measures the percentage change in labour costs with a 1% *increase* in sales revenue, whereas the sum of the coefficients $ß_1 + ß_2$ measures the percentage change in labour costs with a 1% *decrease* in sales revenue. If labour costs display sticky behaviour, the change in labour costs for sales increases is

[19] Prior studies state that using ratios and the ln-form provides several benefits, such as increasing the comparability of the variables between firms, reducing potential bias from heteroscedasticity, decreasing the model's sensitivity towards outliers and enabling an economically meaningful interpretation of the estimated coefficients as percentage changes (M. C. Anderson et al., 2003; Ibrahim et al., 2022).

[20] Additionally including the standalone control variables, i.e. estimating the fully interacted model, yields similar results (untabulated).

greater than the change in labour costs for sales decreases, implying that $ß_2 < 0$, conditional on $ß_1 > 0$ (M. C. Anderson et al., 2003). Consequently, if $ß_2 > 0$ ($ß_2 = 0$), labour costs would behave anti-sticky (symmetrical). The model includes the following firm- and country-level variables based on prior literature to control for economic determinants of asymmetric cost behaviour (M. C. Anderson et al., 2003; Cannon et al., 2020). $SUC_DEC_{i,t}$ is an indicator variable taking the value of one if sales of firm i decrease from year t-2 to year t-1 and from year t-1 to year t, i.e. in two consecutive periods. $GDP_{n,t}$ measures the percentage growth in GDP of country n during year t. Both $SUC_DEC_{i,t}$ and $GDP_{n,t}$ are employed as proxies for managerial expectations about future demand. To proxy the magnitude of adjustment costs, asset intensity ($AINT_{i,t}$) and employee intensity($EINT_{i,t}$),measured as the ln of the ratio of assets or employees to sales of firm i in year t, are included as control variables. $ß_0$ is the constant and $\varepsilon_{i,t}$ represents the error term. *Appendix 3* provides detailed variable definitions.

The model is estimated using a random-effects generalized least squares (GLS) regression[21] with robust standard errors clustered at the firm level, to account for the panel data structure and control for heteroscedasticity and autocorrelation. Given the inclusion of interaction terms, the model is tested for multicollinearity. The initial average variance inflation factor (VIF) exceeds the suggested cut-off value of 10 (VIF = 16.21). Mean-centring the continuous control variables (Banker et al., 2013; Cannon et al., 2020; Prabowo et al., 2018) reduces the VIF to 2.00 and thus effectively mitigates multicollinearity. Mean-adjusting additionally facilitates a straightforward interpretation of the main effects, implying that the coefficients on $\Delta \ln SALES_{i,t}$ ($\Delta \ln SALES_{i,t} xDEC_{i,t}$) can be interpreted as the (incremental) change in labour costs for sales increases (decreases) of a firm that has average values of the continuous control variables (Cannon et al., 2020).

Difference-in-Differences model

The global and exogenous nature of the COVID-19-induced demand shock, as well as the broad cross-country variation in subsequent government interventions suggest the suitability of a difference-in-differences research design for investigating the relation between the stringency of government interventions and cost stickiness (Finne et al., 2024; Goodman-Bacon & Marcus, 2020). A DiD design can leverage quasi-experimental variation arising from the assumption of as-if random assignment to treatment and control and yields the advantage of reducing the likelihood that the estimated association is influenced by unobserved confounding factors. Therefore, an expanded model based on the DiD design of Cannon et al. (2020)[22] is used to examine the relation between the stringency of government interventions during COVID-19 and labour cost stickiness. The properties of the M. C. Anderson et al. (2003) model allow appending additional control variables as interaction terms (Ibrahim et al., 2022). Hence, the equation is complemented by the indicator variables $TREAT_n$ and $POST_t$, as well as the interaction term $TREAT_n xPOST_t$. The model includes the main effects, as well as two-, three- and four-way interactions[23]. The following DiD regression specification is used to compare the changes in the degree of labour cost stickiness for firms subject to stringent government interventions from the pre- to the post-COVID-19-period (treatment firms) with the changes in the degree of labour cost asymmetry for firms not subject to stringent government interventions from the pre- to the post-COVID-19-period (control firms):

$$
\begin{aligned}
\Delta \ln LABOUR_{i,t} = {} & \left(ß_0 + ß_1 TREAT_n + ß_2 POST_t \right. \\
& \left. + ß_3 TREAT_n xPOST_t + ß_4 CONTROLS_{i,t\ /\ n,t}\right) \\
& + \left(\gamma_0 + \gamma_1 TREAT_n + \gamma_2 POST_t + \gamma_3 TREAT_n xPOST_t \right. \\
& \left. + \gamma_4 CONTROLS_{i,t\ or\ n,t}\right) xDEC_{i,t} \\
& + \left(\delta_0 + \delta_1 TREAT_n + \delta_2 POST_t + \delta_3 TREAT_n xPOST_t \right. \\
& \left. + \delta_4 CONTROLS_{i,t\ or\ n,t}\right) x\ \Delta \ln SALES_{i,t} \\
& + \left(\theta_0 + \theta_1 TREAT_n + \theta_2 POST_t + \theta_3 TREAT_n xPOST_t \right. \\
& \left. + \theta_4 CONTROLS_{i,t\ or\ n,t}\right) x\ \Delta \ln SALES_{i,t}\ x\ DEC_{i,t} \\
& + \lambda_i + \mu_t + \varepsilon_{i,t}
\end{aligned}
\tag{2}
$$

$TREAT_n$ is an indicator variable that captures the degree of stringency of government interventions during COVID-19. Following an approach similar to the one used by Akter (2020), $TREAT_n$ equals one for observations from countries whose average stringency index over the COVID-19 period exceeds the median of all sample countries' average stringency indices during the same period[24]. The treatment (control) group consequently consists of observations from countries with a comparably high (low) government stringency index. The assignment of firms to treatment versus control is assumed to be exogenous, as firms cannot choose to be treated or not, i.e. being headquartered in a country with higher or lower government stringency index during COVID-19, at least in the short-term. The indicator variable $POST_t$ is equal to one if the observation is from the COVID-19 period, i.e. 2020-2022, and is equal to zero if the observation stems from the pre-COVID-19 period, i.e. 2017-2019. The

[21] Using an ordinary least squares (OLS) regression yields similar outcomes (untabulated).

[22] Cannon et al. (2020) use a DiD analysis to examine the effect of a legislative shock, namely the enactment of M&A laws, on SG&A cost asymmetry.

[23] Cannon et al. (2020) include only $TREAT_n$ and $TREAT_n xPOST_t$, but do not consider the stand-alone variable $POST_t$ to avoid collinearity issues inherent in their setting. This study incorporates $POST_t$ and its corresponding interaction terms for completeness, thereby estimating the fully interacted model.

[24] One could claim that there is no 'true' control group, as all firms are exposed to the shock of government interventions during COVID-19. However, Atanasov and Black (2016) argue that DiD designs comparing strongly affected versus mildly affected firms are also valid.

outbreak of COVID-19 in 2020 thus constitutes also the point in time where the treatment comes into effect, although this represents a simplification of reality. $CONTROLS_{i,t\ /\ n,t}$ is a vector of the same conventional control variables as in Model (1), namely $SUC_DEC_{i,t}$, $GDP_{n,t}$, $AINT_{i,t}$ and $EINT_{i,t}$.

Similar to the previously described Model (1), the coefficient on $\Delta \ln SALES_{i,t}$ (δ_0) captures labour costs' sensitivity towards sales increases, whereas the sum of the coefficient on $\Delta \ln SALES_{i,t}$ and $\Delta \ln SALES_{i,t} x DEC_{i,t}$ ($\delta_0 + \theta_0$) displays labour costs' sensitivity towards sales decreases. θ_0 thus reflects the average degree of labour cost asymmetry in the full sample. The coefficient on $\Delta \ln SALES_{i,t} x TREAT_n$ (δ_1) and the sum of the coefficients on $\Delta \ln SALES_{i,t} x TREAT_n$ and $\Delta \ln SALES_{i,t} x DEC_{i,t} x TREAT_n$ ($\delta_1 + \theta_1$) captures the difference in labour cost sensitivity towards sales increases and decreases, respectively, between the treatment and the control group. θ_1 thus reflects the average difference in the degree of labour cost asymmetry between firms headquartered in countries which initiated less versus more stringent government interventions during COVID-19. The coefficient on $\Delta \ln SALES_{i,t} x POST_t$ (δ_2) and the sum of the coefficients on $\Delta \ln SALES_{i,t} x POST_t$ and $\Delta \ln SALES_{i,t} x DEC_{i,t} POST_t$ ($\delta_2 + \theta_2$) measures the difference in labour costs' sensitivity towards sales increase and decreases, respectively, between the pre-treatment and the treatment period. θ_2 thus reflects the average difference in the degree of labour cost asymmetry between the pre-COVID-19 and the COVID-19 period. Finally, the coefficient on $\Delta \ln SALES_{i,t} x POST_t x TREAT_n$ (δ_3) and sum of the coefficients $\Delta \ln SALES_{i,t} x POST_t x TREAT_n$ (δ_3)and $\Delta \ln SALES_{i,t} x DEC_{i,t} x POST_t x TREAT_n$ ($\delta_3 + \theta_3$) captures the difference in changes in labour costs' sensitivity towards sales increases and decreases, respectively, from the pre-COVID-19 to the COVID-19 period, between the treatment and the control group. θ_3 thus represents the construct of interest. It displays the difference in the changes in labour cost stickiness from before to after the treatment year 2020 in firms headquartered in countries with stringent government interventions to the changes in labour cost stickiness from before to after the treatment year 2020 in firms located in countries with less stringent government interventions. As Hypothesis 2 is stated in a neutral form, there is no prediction on the sign of θ_3. If θ_3 is positive (negative), there is a negative (positive) relation between the stringency of government interventions during COVID-19 and labour cost stickiness.

Following Cannon et al. (2020), a fixed effects model including firm- (λ_i) and year-fixed effects (μ_t)[25] is used to control for unobservable time-invariant firm characteristics and temporal factors. Robust standard errors are clustered at the country level[26] to reduce heteroscedasticity and autocorrelation. As in Model (1), the initial average VIF exceeds the threshold of 10 (VIF = 23.15) which indicates multicollinearity issues. After mean-centering the continuous control variables (Cannon et al., 2020), the VIF equals 8.80. Thus, the mean-centered variables are used in the following regression analysis.

The validity of a DiD design depends on the assumption of comparability of the treatment and the control group (Goodman-Bacon & Marcus, 2020) in the pre-treatment period. In other words, it depends on the fulfilment of the parallel pre-trends assumption, which requires that the difference in labour cost stickiness between firms in the treatment and the control group would remain constant in the absence of the treatment, i.e. stringent government interventions during COVID-19. To check this key assumption, a labour cost stickiness variable ($STICKY_t$) is created for each year for the treatment and the control group, respectively, and displayed graphically. The most common way to define a firm-level cost stickiness measure in the cost asymmetry literature is the approach of Weiss (2010). The author models $STICKY_t$ as the difference between the ratio of cost changes to sales changes of the most recent period with decreasing sales and the corresponding ratio of cost changes to sales changes of the most recent period with increasing sales, across the last four periods. As the sample of this study spans only six periods, i.e. years, in total, following the approach of Weiss (2010) is not feasible due to significant data loss. Thus, to create $STICKY_t$, this study uses an alternative, yet less common, model based on Zhang et al. (2022). Unlike Zhang et al. (2022) who do not include the main effect of $DEC_{i,t}$ and the two-way interaction term $\Delta \ln SALES_{i,t} x SUC_DEC_{i,t}$, this study utilizes the following fully interacted model:

$$\begin{aligned} \Delta \ln LABOUR_{i,t} = {} & ß_0 + ß_1 \Delta \ln SALES_{i,t} + ß_2 DEC_{i,t} \\ & + ß_3 SUC_{i,t} + ß_4 \Delta \ln SALES_{i,t} x DEC_{i,t} \\ & + ß_5 \Delta \ln SALES_{i,t} x SUC_{i,t} \\ & + ß_6 \Delta \ln SALES_{i,t} x DEC_{i,t} x SUC_{i,t} + \varepsilon_{i,t} \end{aligned} \tag{3}$$

Consistent with Zhang et al. (2022), the model does not entail the control variables used by M. C. Anderson et al. (2003) with the exception of $SUC_DEC_{i,t}$ and defines $STICKY_t$ as the ratio of the coefficient on $\Delta \ln SALES_{i,t} x DEC_{i,t}$ ($ß_4$) to the coefficient on $\Delta \ln SALES_{i,t}$ ($ß_1$)[27]. For a more intuitive interpretation, the ratio of $ß_4/ß_1$ is multiplied by minus one (Zhang et al., 2022). *Figure 2* indicates visually that the parallel pre-trends assumption can be confirmed[28].

The trends in labour cost stickiness between 2017 and 2019, i.e. in the pre-treatment period, are relatively paral-

[25] Including country-fixed effects likely results in removing variance related to the variable of interest $\Delta \ln SALES_{i,t} x DEC_{i,t} x POST_t x TREAT_n$, which is why this study refrains from doing so in the main analysis. It should be noted that omitting country-fixed effects raises the concern that treatment and control group differ in unobserved country-level characteristics which could confound the results. As expected, if country-fixed, industry-fixed, and year-fixed effects are included in the model as a robustness check like suggested by Cannon et al. (2020), the coefficient on $\Delta \ln SALES_{i,t} x DEC_{i,t} x POST_t x TREAT_n$ gets smaller ($\theta_3 = -0.153$, $p < 0.01$) and is only statistically significant in a regression specification without robust and clustered standard errors (untabulated).

[26] Unreported results confirm that the findings are robust to alternatively clustering standard errors at the firm-level.

[27] Zhang et al. (2022) argue that defining $STICKY_{i,t}$ as $ß_4/ß_1$ rather than $ß_4$ on its own is beneficial because the ratio also accounts for labour costs' sensitivity to activity changes, i.e. cost elasticity, in general. However,

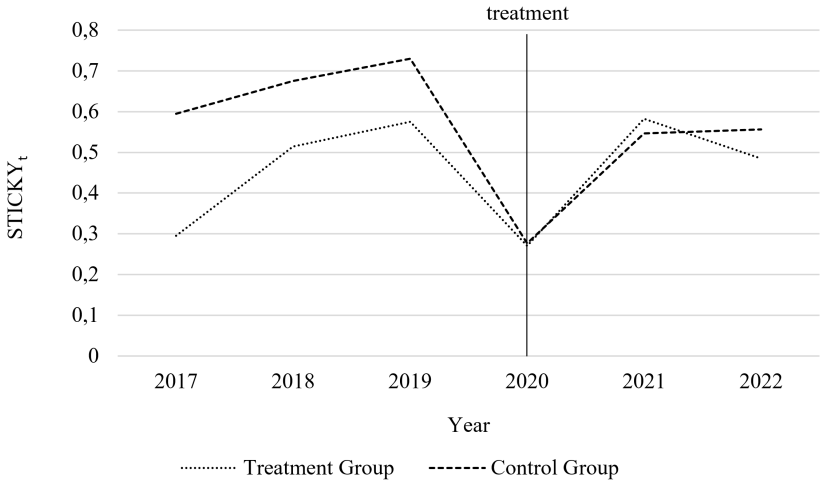


Figure 2: Graphical representation of the parallel pre-trends assumption

lel in the treatment and the control group until 2020. The graph indicates that the treatment group exhibits lower levels of labour cost stickiness compared to the control group before the outbreak of COVID-19 in 2020. With the outbreak of the pandemic, labour cost stickiness in both groups drops significantly. However, during the period from 2020 to 2021, the treatment group displays higher levels of cost stickiness, before falling below the cost stickiness level of the control group after 2021 once again.

Sample split

To test Hypothesis 2 and examine the potential moderating role of national culture in the relation between government interventions during COVID-19 and labour cost stickiness, Model (2) is re-estimated for various subsamples. Kitching et al. (2016) highlight that the six dimensions of national culture are highly correlated, which is also the case in this study's sample, as can be seen in *Table 3*. Thus, to avoid multicollinearity issues, the cultural dimensions are examined separately and not in a single model. Furthermore, including the cultural dimensions as control variables would involve using five-way-interaction terms, which complicates the interpretation of the coefficients (C. X. Chen et al., 2012). Therefore, based on prior research which uses the DiD methodology and examines moderation effects (e.g. Cannon et al., 2020; L. Chen et al., 2022), the sample is divided into two subsamples for each cultural dimension, using the dimensions' median. Observations with values greater than or equal to the median represent the "high"-subsample, while observations with values less than the median make up the "low"-subsample. An indicator variable is used to assign the observations to either group. For example, PDI_high_n (PDI_low_n) is coded one if country n's power distance score is above (below) the median value of power distance in the full sample, and zero otherwise. Subsequently, the DiD Model (2) is re-estimated for both groups in each of the six dimensions of national culture, namely PDI, IDV, MAS, UAI, LTO and IVR. To test Hypothesis 2, it is examined whether the coefficient on $\Delta \ln SALES_{i,t} xDEC_{i,t} xPOST_t xTREAT_n$ (θ_3) varies between the high- and the low-subsample. To determine statistical significance, a z-test based on Clogg et al. (1995)[29]is used. If the difference between the coefficients $\theta_{3,high}$ and $\theta_{3,low}$ is statistically significant, this is an indication for national culture moderating the relation between the stringency of government interventions during COVID-19 and labour cost stickiness.

4. Results and discussion

4.1. Descriptive statistics and correlation results

Descriptive statistics

Panels A and B of *Table 2* provide the descriptive statistics for the main firm- and country-level variables[30], including the respective mean, standard deviation, median and 25%- and 75%-percentiles.

Overall, the descriptive statistics are broadly similar to those of prior studies. However, since there is no study examining a current-period European sample, direct comparisons are hardly feasible. The mean value of labour costs

defining $STICKY_{i,t}$ as ß$_4$ also yields similarly parallel pre-trends.

[28] Though, checking for covariate balance in the pre-treatment period reveals that the treatment and control groups' means of almost all variables differ statistically significantly, as can be seen in *Appendix 4*. By including the corresponding variables in the regression model, the baseline differences are adjusted for. However, even though the pre-treatment trends seem reasonably parallel, the covariate imbalance in the pre-treatment period might raise the concern that treatment and control might as well differ in unobservable characteristics, which would confound the results (Atanasov & Black, 2016).

[29] The z-Test following Clogg et al. (1995) is generally calculated as follows: $z = \frac{\widehat{ß}_1 - \widehat{ß}_2}{\sqrt{[s^2(\widehat{ß}_1) + s^2(\widehat{ß}_2)]}}$.

[30] For variable definitions see *Appendix 3* The variables are presented without being mean-centered.

Table 2: Descriptive statistics

Panel A: Firm-level variables					
Variable	Mean	St. dev.	Q25	Q50	Q75
$\text{LABOUR}_{i,t}$	527.29	1,822.15	13.27	54.82	253.66
$\Delta \ln LABOUR_{i,t}$ [winsorized]	0.08	0.22	-0.02	0.06	0.15
$\text{SALES}_{i,t}$	2,980.00	12,400.00	59.16	251.52	1,273.63
$\Delta \ln SALES_{i,t}$ [winsorized]	0.09	0.31	-0.03	0.07	0.19
$\text{DEC}_{i,t}$	0.31	0.46	0.00	0.00	1.00
$\text{SUC_DEC}_{i,t}$	0.12	0.33	0.00	0.00	0.00
$\text{ASSETS}_{i,t}$	4,420.74	19,700.00	77.09	320.98	1,669.63
$\text{AINT}_{i,t}$ [winsorized]	0.31	0.75	-0.17	0.20	0.66
$\text{EMPLOYEES}_{i,t}$	11,059	39,193	261	1,112	5,150
$\text{EINT}_{i,t}$ [winsorized]	-5.43	0.87	-5.89	-5.37	-4.90
Panel B: Country-level variables					
Variable	Mean	St. dev.	Q25	Q50	Q75
$\text{GDP}_{n,t}$ [winsorized]	1.83	4.32	1.12	2.29	4.35
SI_n	41.95	4.92	37.26	42.90	44.04
TREAT_n	0.66	0.48	0.00	1.00	1.00
POST_t	0.56	0.496	0.00	1.00	1.00
PDI_n	44.42	16.15	35.00	35.00	60.00
PDI_high_n	0.38	0.49	0.00	0.00	1.00
IDV_n	70.70	12.93	67.00	71.00	76.00
IDV_high_n	0.82	0.39	1.00	1.00	1.00
MAS_n	50.27	22.37	42.00	64.00	66.00
MAS_high_n	0.76	0.43	1.00	1.00	1.00
UAI_n	61.62	23.33	35.00	65.00	86.00
UAI_high_n	0.40	0.49	0.00	0.00	1.00
LTO_n	57.86	15.24	51.13	52.90	63.98
LTO_high_n	0.48	0.50	0.00	0.00	1.00
IVR_n	52.70	16.72	40.40	48.55	69.42
IVR_high_n	0.63	0.48	0.00	1.00	1.00
CASES_n	3,897.02	1,210.69	2,724.07	3,600.83	4,861.95
$\ln CASES_n$ [winsorized]	8.21	0.35	7.91	8.19	8.49
DEATHS_n	24.69	7.54	19.36	23.37	31.62
$\ln DEATHS_n$ [winsorized]	3.15	0.34	2.96	3.15	3.45
ES_n	48.96	13.40	35.15	56.14	56.80
$\ln ES_n$ [winsorized]	3.85	0.29	3.56	4.03	4.04

($LABOUR_{i,t}$)is approximately 527.29 million euro. This corresponds to 17.7% of total sales revenue ($SALES_{i,t}$), which amounts to 2,980.00 million euro, on average. The proportion of labour costs to sales is close to the one reported by Hassanein and Younis (2020), who present a ratio of 18.7% for chemical firms in the UK during the 2008 financial

crisis. The variables in monetary units are comparable to BenYoussef et al. (2023), who report average SG&A costs of 471.12 million dollar and average sales revenues of 2,669.08 million dollar for North American companies from 1999 to 2021. All variables reported in euro are heavily skewed to the right. However, this is again consistent with the descriptives of BenYoussef et al. (2023) and constitutes no issue as the ln-specification largely diminishes the skewedness. The average annual growth rate of ln-labour costs is 0.08 and the average annual growth rate of ln-sales is 0.09, which is comparable to the values reported by Cannon et al. (2020), who examine cost asymmetry in an international sample spanning the years 1984 to 2011[31]. $DEC_{i,t}$ has a mean value of 0.31, indicating that 31% of the firms in the sample experience a decrease in sales. Similarly, the mean value of 0.12 for $SUC_DEC_{i,t}$ shows that 12% of the sample firms experience a decrease in sales for two consecutive periods. This is roughly consistent with the descriptive statistics of Prabowo et al. (2018), whose sample entails firms from 22 European countries between 1993 and 2012. The firms included in the sample have on average 4,420.74 million euro in total assets ($ASSETS_{i,t}$) and employ 11,059 employees ($EMPLOYEES_{i,t}$), which is roughly equal to the sample firms in BenYoussef et al. (2023). The asset intensity ($AINT_{i,t}$) in this study is higher than in Cannon et al. (2020), whereas the employee intensity ($EINT_{i,t}$) is lower. This could, for example, reflect the high proportion of manufacturing firms in this study's sample, which exhibit high levels of fixed assets (Subramaniam & Weidenmier Watson, 2016) compared to service firms. Or it could indicate a shift toward increased automatization in the more recent sample, resulting in an increase in asset intensity and a decrease in employee intensity. The average annual GDP growth ($GDP_{n,t}$) is 1.83 percent, suggesting that even when considering the COVID-19 pandemic, the economic conditions in Europe in the sample period are generally positive. The countries in the sample have an average stringency index (SI_n) of 41.95, which is lower than the average SI of 62.66 for North America (BenYoussef et al., 2023). The mean of $TREAT_n$ indicates that 66% of the firms are in the treatment group, whereas the mean of $POST_t$ shows that 56% of the observations are from the COVID-19 period. This implies that the sample includes an adequately balanced number of observations from both treated and untreated firms, as well as from before and after the treatment, i.e. the year 2020. The statistics of the national culture dimensions are broadly consistent with Kitching et al. (2016) who examine an international sample in the period from 1990 to 2013[32].

Correlation results

Panel A of *Table 3* present the results of the Pearson correlation analysis, showing the relationships among the key variables from Model (1). Furthermore, it displays the correlations between the main variables and indicator variables for the years 2017 to 2020 in order to examine changes over time. Since the associations between the independent variables of interest and labour cost stickiness are not captured by standalone variables, but two-, three- and four-way interaction terms, no conclusions on Hypotheses 1 and 2 can be drawn from the correlation results.

As expected, $\Delta \ln LABOUR_{i,t}$ and $\Delta \ln SALES_{i,t}$ are significantly positively correlated (0.599, $p < 0.01$), indicating that a positive change in sales revenue is associated with a positive change in labour costs. $DEC_{i,t}$ and $SUC_DEC_{i,t}$ are negatively correlated to $\Delta \ln LABOUR_{i,t}$ (−0.416, $p < 0.01$; −0.282, $p < 0.01$), suggesting that a (successive) decrease in sales is negatively associated with labour cost growth. The correlation between some variables, such as $DEC_{i,t}$ and $\Delta \ln SALES_{i,t}$ is moderate to high (−0.602, $p < 0.01$) which implies that multicollinearity should be controlled for to avoid issues related to it. The year indicators of the pre-COVID-19 period, i.e. $YEAR_2017_t$ $YEAR_2018_t$ and $YEAR_2019_t$, are not or only weakly correlated to $\Delta \ln LABOUR_{i,t}$, $\Delta \ln SALES_{i,t}$ and $GDP_{n,t}$. However, the indicator variable for the COVID-19 outbreak year 2020 ($YEAR_2020_t$), shows a moderate to strong negative correlation to those three variables and a positive correlation to $DEC_{i,t}$ and $SUC_DEC_{i,t}$. This might be an indication for the adverse economic effects induced by the COVID-19 pandemic, being reflected in declining average sales as well as falling GDP growth rates. Interestingly, the indicators for the years 2021 ($YEAR_2021_t$) and 2022 ($YEAR_2022_t$) depict a positive correlation with $\Delta \ln LABOUR_{i,t}$, $\Delta \ln SALES_{i,t}$, and $GDP_{n,t}$, suggesting that the overall economic condition rapidly resurged after the drop in 2020.

Panel B of *Table 3* shows the associations between $TREAT_n$ and the various dimensions of national culture. All culture variables, except for long-term orientation (LTO_n) and individualism (IDV_n), are significantly correlated with each other, as stated by Kitching et al. (2016). It should be noted that all dimensions of national culture are significantly correlated to $TREAT_n$. This indicates an association between the various cultural dimensions and the likelihood of a country being in the treatment group, meaning its government imposing interventions more stringent than the median of all European countries' means in the sample. For example, as the correlation between LTO_n and $TREAT_n$ is positive and significant (0.328, $p < 0.01$), a government in a long-term oriented country is more likely to impose stringent interventions during COVID-19. In contrast, because the cor-

[31] However, Cannon et al. (2020) examine SG&A cost stickiness instead of labour cost stickiness.

[32] It should be noted that the means of the indicator variables for the cultural dimensions, for example PDI_high_n, differ from 0.5. This suggests that the number of observations with a PDI index higher than the median is unequal to the number of observations with a PDI index lower than the median. This disparity arises as some countries are more heavily represented than others, as can be seen in *Appendix 2*. Although restricting the sample to countries with a similar number of observations would mitigate this imbalance, it would also result in significant data loss. Therefore, this study opts to retain the full dataset, acknowledging the unequal subsample sizes as a potential caveat.

relation between IVR_n is negative and significant (−0.149, $p < 0.01$), a government in a country high in indulgence is less likely to impose stringent interventions. Thus, there is not just a potential moderating effect of national culture on the relation between stringency of government interventions and labour cost stickiness, but likely also a direct relation between national culture and governmental stringency during COVID-19. However, this is not the focus of this study and therefore not examined in further detail.

4.2. Analysis and discussion of the regression results

Basic model

In a first step, Model (1) is employed to assess whether labour costs in Europe are on average sticky, using the full sample which includes the years 2017 to 2022. Columns (I) and (II) of *Table 4* present the regression results for the basic Model (1) with and without controls, respectively.

The results in column (I) show that labour costs in the sample behave sticky on average, as ß$_1$ is positive and statistically significant (ß$_1 = 0.431$, $p < 0.01$) and ß$_2$ is negative and statistically significant (ß$_2 = -0.110$, $p < 0.01$). This implies that labour costs increase by approximately 0.431% per 1% increase in sales and decrease by approximately (0.431 - 0.110 =) 0.321% per 1% decrease in sales. Moreover, the results are robust to controlling for determinants of cost stickiness used in prior studies, as shown in column (II). The significantly positive value of ß$_1$ and the significantly negative value of ß$_2$ continue to exist (ß$_1 = 0.489$, $p < 0.01$; ß$_2 = -0.084$, $p < 0.01$). However, the economic significance of the results could be questioned. Given the average sales of 2,980.00 million euro and the average labour costs of 527.29 million euro in the sample, a 29.8 million euro increase in sales (1% times 2,980.00 million euro) increases labour costs by 2.57 million euro (0.489% times 527.29 million euro), whereas a 29.8 million euro decrease in sales decreases labour costs by only 2.13 million euro ((0.489 - 0.084 =) 0.405% times 527.29 million euro), assuming that the company under consideration has average values for the continuous control variables. The difference of (2.57 – 2.13 =) 0.44 million euro between positive and negative labour cost changes in reaction to sales increases versus sales decreases is rather small, especially in comparison to the overall sales and labour cost figures. The potential lack of economic significance may be related to the magnitude of the coefficients being smaller than in prior studies. For example, Prabowo et al. (2018), who examine labour cost stickiness in a European sample with a similar model, report ß$_1 = 0.673$ and ß$_2 = -0.277$. The smaller magnitude of the coefficients could be attributed to the fact that this study's sample contains the COVID-19 pandemic, which prior research has found to decrease cost stickiness due to rising managerial pessimism about future demand.

The signs and magnitude of the control variables are largely consistent with theory and results of prior studies (e.g. M. C. Anderson et al., 2003; Prabowo et al., 2018). The estimated coefficient on $\Delta \ln SALES_{i,t} x DEC_{i,t} x SUC_DEC_{i,t}$ is positive and significant (ß$_4 = 0.090$, $p < 0.01$), indicating that labour cost stickiness decreases when corporations experience declining sales in two consecutive periods. This is consistent with theory stating that managerial future expectations worsen if a sales drop continues to exist in the subsequent year. The coefficients on the interaction terms $\Delta \ln SALES_{i,t} x DEC_{i,t} x GDP_{n,t}$ and $\Delta \ln SALES_{i,t} x DEC_{i,t} x AINT_{i,t}$ are negative and significant (ß$_6 = -0.005$, $p < 0.05$; ß$_8 = -0.136$, $p < 0.01$), implying that cost stickiness increases in economic growth and asset intensity. Again, this is in line with previous research arguing that macroeconomic growth increases managers' optimism about future demand conditions and asset intensity rises adjustment costs, both of which result in higher levels of cost stickiness. In contrast to M. C. Anderson et al. (2003), this study finds a significantly positive coefficient on $\Delta \ln SALES_{i,t} x DEC_{i,t} x EINT_{i,t}$ (ß$_{10} = 0.044$, $p < 0.05$). However, the fact that the sample period covers the COVID-19 pandemic might explain this result. Banker et al. (2020) claim that during an economic downturn, employment opportunities deteriorate, resulting in an oversupply of labour. This leaves firms in a good position to negotiate and allows them to increasingly rely on temporary workers, for which downward adjustments are less costly. Following the argumentation of C. X. Chen et al. (2012), corporations with greater labour intensity also employ a larger percentage of temporary workers, which might lead employee intensity to be negatively related to cost stickiness.

Although not shown in the tables, the inclusion of firm- and year- as well as country-, industry- and year-fixed effects to control for potentially unobserved factors yields similar results[33]. The overall R-squared of Model (1) including control variables is similar to the adjusted R-squared reported by M. C. Anderson et al. (2003) and indicates that the model explains about 39.7% of the variance in labour cost changes.

To provide insights on whether there is a difference in labour cost stickiness before and during the COVID-19 pandemic and establish a starting point for further analysis, Model (1) is re-estimated for the pre-pandemic period and the pandemic period separately. Whereas observations from the years 2017 to 2019 constitute the pre-COVID-19 subsample, observations from the years 2020 to 2022 make up the COVID-19 subsample. This approach is based on prior literature that examines the influence of COVID-19 on cost stickiness (e.g. Liu and Jin, 2023). The regression results are reported in *Table 5*.

Column (I) shows the results for the sample which is restricted to the years prior to COVID-19, i.e. 2017-2019. They are qualitatively the same as the results reported for the full sample in *Table 4*. Most importantly, ß$_1$ is positive and sta-

[33] Although the results of a Hausman (1978) test indicate that the fixed-effects model is more appropriate, this study reports the random effects model for consistency with M. C. Anderson et al. (2003). They used a random coefficient specification to test their regression model and report the fixed effects model only as a robustness check.

Table 3: Correlation results

Panel A: Main variables related to Model (1)

	$\Delta \ln LABOUR_{i,t}$	$\Delta \ln SALES_{i,t}$	$DEC_{i,t}$	$SUC_DEC_{i,t}$	$GDP_{n,t}$	$AINT_{i,t}$	$EINT_{i,t}$	$YEAR_2017_t$	$YEAR_2018_t$	$YEAR_2019_t$	$YEAR_2020_t$	$YEAR_2021_t$	$YEAR_2022_t$
$\Delta \ln LABOUR_{i,t}$	1.000												
$\Delta \ln SALES_{i,t}$	0.599*	1.000											
$DEC_{i,t}$	-0.416*	-0.602*	1.000										
$SUC_DEC_{i,t}$	-0.281*	-0.345*	0.560*	1.000									
$GDP_{n,t}$	0.125*	0.190*	-0.209*	-0.090*	1.000								
$AINT_{i,t}$	-0.014	-0.178*	0.115*	0.089*	-0.024*								
$EINT_{i,t}$	-0.079*	-0.092*	0.107*	0.060*	-0.023*	0.134*	1.000						
$YEAR_2017_t$	-0.036*	-0.024*	-0.005	-0.024*	0.082*	-0.051*	0.013	1.000					
$YEAR_2018_t$	-0.027*	-0.025*	0.001	0.021*	0.019	-0.046*	0.012	-0.161*	1.000				
$YEAR_2019_t$	-0.005	-0.040*	0.010	0.006	0.000	-0.022*	0.012	-0.171*	-0.182*	1.000			
$YEAR_2020_t$	-0.137*	-0.187*	0.247*	0.111*	-0.807*	0.053*	0.039*	-0.179*	-0.190*	-0.203*	1.000		
$YEAR_2021_t$	0.029*	0.081*	-0.068*	0.006	0.511*	0.053*	-0.001	-0.188*	-0.199*	-0.213*	-0.222*	1.000	
$YEAR_2022_t$	0.162*	0.179*	-0.175*	-0.115*	0.182*	0.001	-0.068*	-0.193*	-0.204*	-0.218*	-0.228*	-0.240*	1.000

Panel B: Main variables related to Model (2)

	$TREAT_SI_n$	PDI_n	IDV_n	MAS_n	UAI_n	LTO_n	IVR_n
$TREAT_SI_n$	1.000						
PDI_n	0.116*	1.000					
IDV_n	0.263*	-0.422*	1.000				
MAS_n	0.498*	0.074*	0.212*	1.000			
UAI_n	0.108*	0.816*	-0.654*	0.221*	1.000		
LTO_n	0.328*	0.025*	-0.011	0.301*	0.191*	1.000	
IVR_n	-0.149*	-0.623*	0.541*	-0.318*	-0.804*	-0.251*	1.000

* indicates two-tailed significance at the 1 percent level.

Table 4: Regression results Model (1) – basic model

Variables	(I) Without controls	(II) With controls
$\Delta \ln SALES_{i,t}$ ($ß_1$)	**0.431*****	**0.489*****
	(0.014)	(0.014)
$\Delta \ln SALES_{i,t} xDEC_{i,t}$ ($ß_2$)	**-0.110*****	**-0.084*****
	(0.023)	(0.032)
$\Delta \ln SALES_{i,t} xSUC_DEC_{i,t}$ ($ß_3$)		*omitted*
		(.)
$\Delta \ln SALES_{i,t} xDEC_{i,t}\, xSUC_DEC_{i,t}$($ß_4$)		**0.090*****
		(0.030)
$\Delta \ln SALES_{i,t} xGDP_{n,t}$ ($ß_5$)		-0.001
		(0.002)
$\Delta \ln SALES_{i,t} xDEC_{i,t}\, xGDP_{n,t}$($ß_6$)		**-0.005****
		(0.003)
$\Delta \ln SALES_{i,t} xAINT_{i,t}$ ($ß_7$)		**-0.087*****
		(0.014)
$\Delta \ln SALES_{i,t} xDEC_{i,t} xAINT_{i,t}$ ($ß_8$)		**-0.136*****
		(0.017)
$\Delta \ln SALES_{i,t} xEINT_{i,t}$ ($ß_9$)		**0.083*****
		(0.012)
$\Delta \ln SALES_{i,t} xDEC_{i,t} xEINT_{i,t}$ ($ß_{10}$)		**0.044****
		(0.018)
Constant ($ß_0$)	**0.039*****	**0.037*****
	(0.002)	(0.002)
Number of observations	15,446	15,446
R^2 (overall)	0.364	0.397

Robust standard errors clustered at the firm-level are reported in parentheses.
*, **, *** indicate two-tailed significance at the 10 percent, 5 percent, and 1 percent levels.

tistically significant ($ß_1 = 0.559$, $p < 0.01$) and $ß_2$ is negative and statistically significant ($ß_2 = -0.205$, $p < 0.01$), which implies that labour costs in the pre-COVID-19 period are on average sticky. The magnitude of both $ß_1$ and $ß_2$ is larger in the pre-COVID-19 sample than in the full sample and now more comparable to the coefficients reported in prior research (e.g. Prabowo et al., 2018). This confirms the previously made supposition that the full samples' coefficients on $\Delta \ln SALES_{i,t}$ and $\Delta \ln SALES_{i,t} xDEC_{i,t}$ are smaller than those reported in previous studies due to the inclusion of the COVID-19 crisis in the sample period. Now economic significance is more likely given as well. Assuming average sales of 2,980.00 million euro and average labour costs of 527.29 million euro, a 29.8 million euro increase in sales (1% times 2,980.00 million euro) increases labour costs by 2.95 million euro (0.559% times 527.29 million euro), whereas a 29.8 million euro decrease in sales decreases labour costs by only 1.87 million euro ((0.559 - 0.205 =) 0.354% times 527.29 million euro), assuming that the company under consideration has average values for the continuous control variables.

Column (II) reports the results for the sample which is restricted to the COVID-19 years, i.e. 2020 to 2022. The coefficient on $\Delta \ln SALES_{i,t}$ ($ß_1$) remains positive and statistically significant ($ß_1 = 0.482$, $p < 0.01$). Interestingly, its size is smaller compared to the pre-COVID-19 subsample, which indicates a decline in firms' cost elasticity for sales increases during the pandemic. A pairwise comparison of the coefficients of interest using a z-test based on Clogg et al. (1995) confirms the statistical significance of the difference between the coefficients (2.502, $p < 0.05$). This finding contrasts

Table 5: Regression results pre-COVID-19 versus COVID-19 period

Variables	(I) Pre-COVID-19 period	(II) COVID-19 period
$\Delta \ln SALES_{i,t}$ ($ß_1$)	**0.559*****	**0.482*****
	(0.026)	(0.017)
$\Delta \ln SALES_{i,t} xDEC_{i,t}$ ($ß_2$)	**-0.205*****	**-0.054**
	(0.057)	(0.039)
$\Delta \ln SALES_{i,t} xSUC_DEC_{i,t}$ ($ß_3$)	*omitted*	*omitted*
	(.)	*(.)*
$\Delta \ln SALES_{i,t} xDEC_{i,t} xSUC_DEC_{i,t}$ ($ß_4$)	**0.126****	**0.076****
	(0.061)	(0.035)
$\Delta \ln SALES_{i,t} xGDP_{n,t}$ ($ß_5$)	0.000	-0.001
	(0.016)	(0.002)
$\Delta \ln SALES_{i,t} xDEC_{i,t} xGDP_{n,t}$ ($ß_6$)	-0.020	-0.004
	(0.025)	(0.003)
$\Delta \ln SALES_{i,t} xAINT_{i,t}$ ($ß_7$)	**-0.134*****	**-0.068*****
	(0.025)	(0.015)
$\Delta \ln SALES_{i,t} xDEC_{i,t} xAINT_{i,t}$ ($ß_8$)	**-0.148*****	**-0.138*****
	(0.028)	(0.020)
$\Delta \ln SALES_{i,t} xEINT_{i,t}$ ($ß_9$)	0.137	**0.051*****
	(0.024)	(0.013)
$\Delta \ln SALES_{i,t} xDEC_{i,t} xEINT_{i,t}$ ($ß_{10}$)	0.033	**0.055*****
	(0.031)	(0.021)
Constant ($ß_0$)	**0.026*****	**0.036*****
	(0.003)	(0.003)
Z-Value (difference of $ß_1$ between pre-COVID-19 and COVID-19 subsample)	**2.502****	
Z-Value (difference of $ß_2$ between pre-COVID-19 and COVID-19 subsample)	**-2.185****	
Number of observations	6,786	8,660
R^2 (overall)	0.373	0.413

Robust standard errors clustered at the firm-level are reported in parentheses.
*, **, *** indicate two-tailed significance at the 10 percent, 5 percent, and 1 percent levels.

with previous studies that report increased levels of cost elasticity, interpreted as a lower proportion of fixed to variable costs, during periods of heightened risk and uncertainty (e.g. Holzhacker et al., 2015a). Firms with a cost structure characterized by a high proportion of fixed costs, or in other words lower cost elasticity, are argued to require higher levels of activity to break even, compared to firms with higher cost elasticity (Holzhacker et al., 2015b). Consequently, firms increase their cost elasticity in times of crises, to be in a better position to avoid economic losses. However, unlike studies focusing on cost elasticity in general, this study distinguishes between cost behaviour for sales increases versus decreases. Therefore, the results are not directly comparable. Detached from the definition of cost elasticity as measuring the proportion of fixed to variable costs, the size of the coefficient on $\Delta \ln SALES_{i,t}$ can also be interpreted as reflecting the magnitude of managers' resource expansions in response to sales growth (e.g. Cannon et al., 2020; Ma et al., 2021). In this context, the findings could indicate that the COVID-19 pandemic increases managerial pessimism about future demand

expectations in general, and the uncertainty about the duration of the sales growth in particular. Managers might be more likely to take a cautious 'wait and see'-approach and, for example, delay (re-)hiring employees, even if sales increase. Hence, the observed decrease in cost elasticity when sales rise and the increase of cost elasticity when sales fall (i.e., the decrease in cost stickiness) during the pandemic in this study can be reconciled with the cost elasticity literature's view that firms adjust their resources to improve flexibility when uncertainty is high.

As can be seen in columns (I) and (II) respectively, the coefficient on $\Delta \ln SALES_{i,t} x DEC_{i,t}$ is negative and significant ($ß_2 = -0.205$, $p < 0.01$) in the pre-COVID-19 sample, whereas it is close to zero and statistically insignificant ($ß_2 = -0.054$, $p > 0.1$) in the COVID-19 sample. The statistical significance of the difference between the coefficients in the pre-COVID-19 and the COVID-19 subsample is confirmed by a z-test (-2.185, $p < 0.05$). Hence, during the pandemic, there is no difference in the magnitude of labour cost changes between sales increases and decreases like in the pre-COVID-19 period. Consequently, the COVID-19 subsample shows no indication of labour cost stickiness, on average. This in line with theory and the findings of prior literature (e.g. Ghazy et al., 2024; Kwak et al., 2021) which suggest that the uncertainty during the pandemic increases managers' pessimism about the duration of the sales decline and incentivizes corporate decision-makers to discard unused resources. These results on cost elasticity and cost stickiness before and during the COVID-19 pandemic constitute the starting point for further analyses and interpretations.

Difference-in-differences model

Table 6 presents the regression results for Model (2) which examines the relation between the stringency of government interventions during COVID-19 and labour cost stickiness.

Like in Model (1) the coefficients on $\Delta \ln SALES_{i,t}$ and $\Delta \ln SALES_{i,t} x DEC_{i,t}$ are significantly positive ($\delta_0 = 0.507$, $p < 0.01$) and negative ($\theta_0 = -0.262$, $p < 0.01$), respectively. Hence, the results reconfirm that labour costs in the full sample are on average sticky. This implies that labour costs increase by 0.507% per 1% increase in sales and decrease by ($\delta_0 + \theta 0 = 0.507 - 0.262 =$) 0.245% per 1% decrease in sales, on average.

The coefficient on $TREAT_n x \Delta \ln SALES_{i,t}$ (δ_1) is not statistically significant ($\delta_1 = -0.108$, $p > 0.1$), showing that for sales increases, there is no significant difference in cost elasticity between treatment and control group, on average. The coefficient on $TREAT_n x \Delta \ln SALES_{i,t} x DEC_{i,t}$, is positive and statistically significant ($\theta_1 = 0.191$, $p < 0.1$). This implies that labour cost stickiness is, on average, significantly lower in the treatment group than in the control group over the full sample period, as can be seen in the graphical representation of $STICKY_t$ in *Figure 2* as well. However, the difference in the pre-treatment period is controlled for by utilizing the DiD design, assuming that parallel pre-trends exist. It should be noted that the variable $TREAT_n$ ($ß_1$) is omitted when estimating the model due to the simultaneous inclusion of firm-fixed effects. The effect of $TREAT_n$ is absorbed by the firm-fixed effects because a firm's assignment to the treatment or control group is constant over the sample period, as long as the location of a firm's headquarters remains the same, which is the case in this study. To avoid collinearity issues, $TREAT_n$is automatically dropped (Cannon et al., 2020).

The coefficient on $POST_t x \Delta \ln SALES_{i,t}$ is negative, but insignificant ($\delta_2 = -0.039$, $p > 0.1$), implying that cost elasticity for sales increases during the pandemic is not significantly different than in the pre-COVID-19 period. However, the coefficient on $POST_t x \Delta \ln SALES_{i,t} x DEC_{i,t}$ is positive and statistically significant ($\theta_2 = 0.258$, $p < 0.01$) and hence supports the finding of prior literature that labour cost stickiness decreases with the outbreak of COVID-19. Thus, as of 2020, labour costs increase on average by approximately ($\delta_0 + \delta_2 = 0.507 - 0.039 =$) 0.468% per 1% increase in sales and decrease by approximately ($\delta_0 + \delta_2 + \theta_0 + \theta_2 = 0.507 - 0.039 - 0.262 + 0.258 =$) 0.464% per 1% decrease in sales. Consistent with the results shown in *Table 5*, this highlights that labour cost stickiness is no longer present following the outbreak of COVID-19. During the pandemic, labour costs behave symmetrically for sales increases and sales decreases. This supports the argumentation of prior authors who claim that COVID-19 induces uncertainty that, in turn, increases managerial pessimism about the duration of the decline in sales and is therefore negatively related to cost stickiness.

Hypothesis 1 states that there is no relation between the stringency of government interventions during COVID-19 and labour cost stickiness, as theory predicts opposing directions of the association. However, the coefficient on $\Delta \ln SALES_{i,t} x DEC_{i,t} x TREAT_n x POST_t$ is negative and statistically significant ($\theta_3 = -0.260$, $p < 0.05$), showing that treatment firms exhibit significantly higher levels of labour cost stickiness after the treatment, i.e. the implementation of stringent government interventions during COVID-19, than control firms, when holding everything else equal. Thus, for a 1% decrease in sales after the treatment, treatment firms exhibit a ($|\delta_3 + \theta_3| = |0.117 - 0.260| =$) 0.143 percentage point lower decrease in labour costs than the control firms, after accounting for pre-treatment trends. This indicates that treatment firms' labour costs are less sensitive to negative sales changes after the treatment. Put into an economic context, this means that the difference in labour cost sensitivity to sales decreases after 2020 between treatment and control group is approximately half as large as the average labour cost sensitivity to sales decreases in the full sample (58.37% = 0.143% / 0.245% = (|0.117 - 0.260|) / (|0.507 - 0.262|) = ($|\delta_3 + \theta_3|$) / ($|\delta_0 + \theta_0|$)). Additionally, the DiD coefficient on $\Delta \ln SALES_{i,t} x DEC_{i,t} x TREAT_n x POST_t$ (θ_3) is comparable in magnitude to DiD coefficients in previous studies which are deemed economically significant (e.g. Cannon et al., 2020; L. Chen et al., 2022).

The coefficient on $\Delta \ln SALES_{i,t} x TREAT_n x POST_t$ is positive and statistically significant ($\delta_3 = 0.117$, $p < 0.1$), im-

Table 6: Regression results Model (2) – difference-in-differences model

Variables	
Main effects	
$Constant\,(ß_0)$	**0.047*****
	(0.008)
$TREAT_n\,(ß_1)$	*omitted*
	(.)
$POST_t\,(ß_2)$	**-0.0369***
	(0.020)
$SUC_DEC_{i,t}\,(ß_{4,a})$	-0.010
	(0.008)
$GDP_{n,t}\,(ß_{4,b})$	-0.003
	(0.002)
$AINT_{i,t}\,(ß_{4,c})$	0.014
	(0.013)
$EINT_{i,t}\,(ß_{4,d})$	**0.088*****
	(0.015)
$DEC_{i,t}\,(\gamma_0)$	**-0.021****
	(0.010)
$\Delta \ln SALES_{i,t}\,(\delta_0)$	**0.507*****
	(0.060)
Two-way interaction terms	
$TREAT_n x POST_t\,(ß_3)$	-0.008
	(0.013)
$TREAT_n x DEC_{i,t}\,(\gamma_1)$	-0.003
	(0.016)
$TREAT_n x \Delta \ln SALES_{i,t}\,(\delta_1)$	-0.108
	(0.084)
$POST_t x DEC_{i,t}\,(\gamma_2)$	0.020
	(0.012)
$POST_t x \Delta \ln SALES_{i,t}\,(\delta_2)$	-0.039
	(0.047)
$SUC_{i,t} x DEC_{i,t}\,(\gamma_{4,a})$	*omitted*
	(.)
$GDP_{n,t} x DEC_{i,t}\,(\gamma_{4,b})$	0.000
	(0.001)
$AINT_{i,t} x DEC_{i,t}\,(\gamma_{4,c})$	**0.010****
	(0.004)
$EINT_{i,t} x DEC_{i,t}\,(\gamma_{4,d})$	0.005
	(0.006)
$SUC_DEC_{i,t} x \Delta \ln SALES_{i,t}\,(\delta_{4,a})$	0.067
	(0.043)
$GDP_{n,t} x \Delta \ln SALES_{i,t}\,(\delta_{4,b})$	0.001
	(0.003)
$AINT_{i,t} x \Delta \ln SALES_{i,t}\,(\delta_{4,c})$	**-0.099*****
	(0.015)
$EINT_{i,t} x \Delta \ln SALES_{i,t}\,(\delta_{4,d})$	**0.072*****
	(0.017)
$\Delta \ln SALES_{i,t} x DEC_{i,t}\,(\theta_0)$	**-0.262*****
	(0.073)
Three-way interaction terms	
$TREAT_n x POST_t x DEC_{i,t}\,(\gamma_3)$	-0.004
	(0.014)
$TREAT_n x POST_t x \Delta \ln SALES_{i,t}\,(\delta_3)$	**0.117***
	(0.066)
$TREAT_n x \Delta \ln SALES_{i,t} x DEC_{i,t}\,(\theta_1)$	**0.191***
	(0.096)
$POST_t x \Delta \ln SALES_{i,t} x DEC_{i,t}\,(\theta_2)$	**0.258*****
	(0.081)
$SUC_DEC_{i,t} x \Delta \ln SALES_{i,t} x DEC_{i,t}\,(\theta_{4,a})$	*omitted*
	(.)
$GDP_{n,t} x \Delta \ln SALES_{i,t} x DEC_{i,t}\,(\theta_{4,b})$	-0.005
	(0.003)
$AINT_{i,t} x \Delta \ln SALES_{i,t} x DEC_{i,t}\,(\theta_{4,c})$	0.006
	(0.020)
$EINT_{i,t} x \Delta \ln SALES_{i,t} x DEC_{i,t}\,(\theta_{4,d})$	0.010
	(0.031)
Four-way interaction term	
$\Delta \ln SALES_{i,t} x DEC_{i,t} x TREAT_n x POST_t\,(\theta_3)$	**-0.260****
	(0.119)
Firm-fixed effects	Yes
Year-fixed effects	Yes
Number of observations	15,446
R^2 (overall)	0.284

Robust standard errors clustered at the country-level are reported in parentheses.
*, **, *** indicate two-tailed significance at the 10 percent, 5 percent, and 1 percent levels.

plying that treatment firms, headquartered in countries with a higher stringency index, exhibit significantly higher levels of cost elasticity for sales increases after the treatment than control firms do. Hence, for an 1% increase in sales after the treatment, i.e. after the outbreak of COVID-19, treatment firms exhibit a (δ_3 =) 0.117 percentage point higher increase in labour costs than the control firms, after accounting for pre-treatment trends. This implies that treatment firms'

labour costs are more sensitive to positive sales changes than control firms' labour costs after the treatment.[34]

Summarizing, the negative sign of θ_3 suggests a positive relation between the stringency of government interventions during COVID-19 and labour cost stickiness. Thus, Hypothesis 1 can be rejected. This result is in line with the strand of the finance literature which argues that stringent government interventions signal effective pandemic control to investors. Hence, also corporate decision-makers' perceived uncertainty is argued to be reduced, which leads them to form more optimistic expectations about future demand. Managers are consequently more inclined to expand labour resources, i.e. hire new employees or increase the existing employees' working hours, in response to sales increases as they believe in the permanence of the sales growth. At the same time, managers are less likely to divest labour resources, i.e. fire employees or reduce their working hours when sales decrease, because they expect sales drops to be only temporary. Consequently, the results indicate that stringent government interventions diminish the COVID-19 induced pessimism and the resulting decrease in labour cost stickiness to some extent.

It should be noted that the result contradicts the findings of Ghazy et al. (2024) and BenYoussef et al. (2023), who report that the stringency of government interventions during COVID-19 is positively associated to cost anti-stickiness, or in other words, negatively associated to cost stickiness. However, there are some differences between their research methodologies and the one of this study which might explain the differing results. BenYoussef et al. (2023, p. 14) use a different dependent variable, focusing on abnormal SG&A costs, which they define as the "discrepancy between the actual SG&A cost and an out-of-sample forecast from a benchmark model similar to the framework [...] proposed by M. C. Anderson et al. (2003) [...]", instead of the logarithmic change in labour costs. Ghazy et al. (2024) define the independent variable as strictness of workplace closure, a variable ranging from zero (no measures) to three (closure of all but essential workplace such as hospitals or grocery stores). Furthermore, Ghazy et al. (2024) do not attempt to proxy governmental stringency, but rather overall COVID-19-induced uncertainty. Both studies examine operating costs instead of labour costs. Moreover, they both analyse North American firms, whereas this study focuses on a European sample. Lastly, their sample periods span the years 2020 to 2021 (Ghazy et al., 2024) and 1999 to 2021 (BenYoussef et al., 2023) respectively, which differs from the sample period examined in this study.

Robustness checks

To be considered valid, a DiD design must fulfil the only-through condition (Atanasov & Black, 2016, p. 217), implying that "the shock must be 'isolated' – there must be no other shock, at around the same time, that could also affect treated firms differently than control firms". However, it could be argued that this is not the case in this study's setting. Given that the treatment coincides with the outbreak of the COVID-19 pandemic in 2020, it is likely that additional shocks, which may differently affect the treatment and control group, occur simultaneously. To mitigate this endogeneity concern, two robustness checks are conducted. This study follows Akter (2020) and BenYoussef et al. (2023) in adding COVID-19 cases and COVID-19-related deaths, as well as the degree of governmental economic support during COVID-19 as additional variables[35] to Model (2).

Greater exposure to the pandemic, measured by COVID-19 cases and deaths, might be a confounding factor as it could lead governments to implement stricter interventions (Pulejo & Querubín, 2021) and simultaneously may induce declining cost stickiness resulting from increased managerial pessimism about future demand (Ghazy et al., 2024). Thus, the natural logarithm of country-level cumulated COVID-19 cases as well as COVID-19 related deaths per 10,000 population[36] over the entire COVID-19 period are added as controls (Akter, 2020).

The degree of governmental economic support should also be controlled for, as it is shown to be correlated with the stringency of governmental interventions and it is likely associated with managerial optimism about future demand, as managers might expect governmental economic support to mitigate potential negative effects of stringent measures on the economy (Alfano et al., 2022). Hence, a country's average economic support index over the entire COVID-19 period, measuring the degree of income support, debt and contract relief, fiscal stimuli and international support, is included in Model (2). *Table 7* shows the results of the robustness checks[37].

As can be seen in column (I) and column (II), respectively, the coefficient of interest on the four-way interaction term

[34] The difference-in-difference estimators for cost elasticity and cost stickiness, i.e. δ_3 and θ_3, can also be calculated based on the remaining coefficients (see for example Jiménez and Perdiguero, 2019). They are computed as follows: $\delta_3 = 0.117 = [(\delta_0+\delta_1+\delta_2+\delta_3)-(\delta_0+\delta_1)]-[(\delta_0+\delta_2)-\delta_0)] = [(0.507+(-0.108)+(-0.039)+0.117)--(0.507+(-0.108))]--[(0.507+(-0.039))--0.507]$ and $\theta_3 = -0.260 = [(\theta_0+\theta_1+\theta_2+\theta_3)-(\theta_0+\theta_1)]-[(\theta_0+\theta_2)-\theta_0)] = [((-0.262)+0.191+0.258+(-0.260))-((-0.262)+0.191)]-[((-0.262)+0.258)--(-0.262)]$. Hence, these estimators represent the difference in the change in labour cost sensitivity to sales increases and sales decreases, respectively, from the pre-treatment to the post-treatment period in the treatment group, relative to the corresponding change in the control group.

[35] For consistency, the additional control variables are mean-centred and transformed to the ln-form and are added to the model as main variables as well as two- and three-way interactions.

[36] Data on the number of COVID-19 cases and COVID-19 related deaths is obtained from the OxCGRT database. The population data used to scale cases and deaths is drawn from the World Bank website (https://data.worldbank.org/indicator/SP.POP.TOTL).

[37] The additional control variables $\ln CASES_n$, $\ln DEATHS_n$ and $\ln ES_n$ are defined as cumulated ($\ln CASES_n$ and $\ln DEATHS_n$) or average ($\ln ES_n$) values over the entire COVID-19 period. This is done for consistency with the assignment of firms to the treatment or control group based on the mean value of their headquarter country's stringency index over the same period. However, as untabulated results show, including $\ln CASES_{n,t}$, $\ln DEATHS_{n,t}$ and $\ln ES_{n,t}$ based on yearly values yields similar outcomes.

Table 7: Regression results Model (2) with additional controls

Variables	(I) Including $\ln CASES_n$ and $\ln DEATHS_n$	(II) Including $\ln ES_n$
Main effects		
Constant($ß_0$)	0.054	-0.018
	(0.095)	(0.036)
$TREAT_n$ ($ß_1$)	*omitted*	*omitted*
	(.)	*(.)*
$POST_t$ ($ß_2$)	*omitted*	*omitted*
	(.)	*(.)*
$SUC_DEC_{i,t}$ ($ß_{4,a}$)	-0.010	-0.010
	(0.007)	(0.007)
$GDP_{n,t}$ ($ß_{4,b}$)	**-0.003*****	**-0.003*****
	(0.001)	(0.007)
$AINT_{i,t}$ ($ß_{4,c}$)	0.014	0.014
	(0.012)	(0.012)
$EINT_{i,t}$ ($ß_{4,d}$)	**0.089*****	**0.088*****
	(0.014)	(0.014)
$\ln CASES_n$ ($ß_{4,e}$)	0.003	
	(0.017)	
$\ln DEATHS_n$ ($ß_{4,f}$)	-0.003	
	(0.016)	
$\ln ES_n$ ($ß_{4,g}$)		**-0.030***
		(0.017)
$DEC_{i,t}$ (γ_0)	-0.113	0.063
	(0.124)	(0.051)
$\Delta \ln SALES_{i,t}$ (δ_0)	0.519	**0.793*****
	(0.440)	(0.180)
Two-way interaction terms		
$TREAT_n x POST_t$ ($ß_3$)	-0.008	0.002
	(0.014)	(0.012)
$TREAT_n x DEC_{i,t}$ (γ_1)	-0.003	-0.004
	(0.013)	(0.013)
$TREAT_n x \Delta \ln SALES_{i,t}$ (δ_1)	**-0.109***	**-0.109***
	(0.063)	(0.063)
$POST_t x DEC_{i,t}$ (γ_2)	0.181	-0.120
	(0.214)	(0.086)
$POST_t x \Delta \ln SALES_{i,t}$ (δ_2)	-0.052	**-0.523***
	(0.756)	0.302
$SUC_{i,t} x DEC_{i,t}$ ($\gamma_{4,a}$)	*omitted*	*omitted*
	(.)	*(.)*
$GDP_{n,t} x DEC_{i,t}$ ($\gamma_{4,b}$)	0.000	0.000
	(0.001)	(0.001)
$AINT_{i,t} x DEC_{i,t}$ ($\gamma_{4,c}$)	**0.009****	0.010
	(0.008)	(0.008)
$EINT_{i,t} x DEC_{i,t}$ ($\gamma_{4,d}$)	0.004	0.004
	(0.006)	(0.006)
$\ln CASES_n x DEC_{i,t}$ ($\gamma_{4,e}$)	-0.019	
	(0.022)	
$\ln DEATHS_n x DEC_{i,t}$ ($\gamma_{4,f}$)	-0.002	
	(0.022)	

Table 7 — continued

$\ln ES_n x DEC_{i,t}$ ($\gamma_{4,g}$)		**0.039***
		(0.023)
$SUC_DEC_{i,t} x \Delta \ln SALES_{i,t}$ ($\delta_{4,a}$)	*omitted*	**0.068***
	(.)	(0.039)
$GDP_{n,t} x \Delta \ln SALES_{i,t}$ ($\delta_{4,b}$)	0.001	0.000
	(0.004)	(0.004)
$AINT_{i,t} x \Delta \ln SALES_{i,t}$ ($\delta_{4,c}$)	**-0.099*****	**-0.096*****
	(0.021)	(0.021)
$EINT_{i,t} x \Delta \ln SALES_{i,t}$ ($\delta_{4,d}$)	**0.070*****	**0.071*****
	(0.019)	(0.019)
$\ln CASES_n x \Delta \ln SALES_{i,t}$ ($\delta_{4,e}$)	-0.019	
	(0.080)	
$\ln DEATHS_n x \Delta \ln SALES_{i,t}$ ($\delta_{4,f}$)	0.054	
	(0.072)	
$\ln ES_n x \Delta \ln SALES_{i,t}$ ($\delta_{4,g}$)		**0.133***
		(0.080)
$\Delta \ln SALES_{i,t} x DEC_{i,t}$ (θ_0)	0.329	**-0.513****
	(0.651)	(0.262)
Three-way interaction terms		
$TREAT_n x POST_t x DEC_{i,t}$ (γ_3)	0.002	-0.016
	(0.021)	(0.019)
$TREAT_n x POST_t x \Delta \ln SALES_{i,t}$ (δ_3)	0.107	0.071
	(0.082)	(0. 076)
$TREAT_n x \Delta \ln SALES_{i,t} x DEC_{i,t}$ (θ_1)	**0.194***	**0.191***
	(0.109)	(0.108)
$POST_t x \Delta \ln SALES_{i,t} x DEC_{i,t}$ (θ_2)	-0.763	0.682
	(1.119)	(0.439)
$SUC_DEC_{i,t} x \Delta \ln SALES_{i,t} x DEC_{i,t}$ ($\theta_{4,a}$)	**0.066***	*omitted*
	(0.038)	*(.)*
$GDP_{n,t} x \Delta \ln SALES_{i,t} x DEC_{i,t}$ ($\theta_{4,b}$)	-0.004	-0.004
	(0.005)	(0.005)
$AINT_{i,t} x \Delta \ln SALES_{i,t} x DEC_{i,t}$ ($\theta_{4,c}$)	0.004	0.004
	(0.032)	(0.032)
$EINT_{i,t} x \Delta \ln SALES_{i,t} x DEC_{i,t}$ ($\theta_{4,d}$)	0.013	0.012
	(0.032)	(0.030)
$\ln CASES_n x \Delta \ln SALES_{i,t} x DEC_{i,t}$ ($\theta_{4,e}$)	0.143	
	(0.116)	
$\ln DEATHS_n x \Delta \ln SALES_{i,t} x DEC_{i,t}$ ($\theta_{4,f}$)	-0.040	
	(0.113)	
$\ln ES_n x \Delta \ln SALES_{i,t} x DEC_{i,t}$ ($\theta_{4,g}$)		-0.116
		(0.116)
Four-way interaction term		
$\Delta \ln SALES_{i,t} x DEC_{i,t} x TREAT_n x POST_t$ (θ_3)	**-0.306****	**-0.221***
	(0.135)	(0.120)
Firm-fixed effects	Yes	Yes
Year-fixed effects	Yes	Yes
Number of observations	15,446	15,446
R^2 (overall)	0.283	0.285

Robust standard errors clustered at the firm-level are reported in parentheses.
*, **, *** indicate two-tailed significance at the 10 percent, 5 percent, and 1 percent levels.

$\Delta \ln SALES_{i,t} xDEC_{i,t} xTREAT_n xPOST_t$ remains negative and significant after controlling for pandemic affectedness ($\theta_3 = -0.306$, $p < 0.05$) and governmental economic support ($\theta_3 = -0.221$, $p < 0.1$). These results indicate that the positive association between stringent government interventions during COVID-19 and labour cost stickiness continues to exist after controlling for factors that change simultaneously with the treatment and hence could constitute alternative explanations for the change in labour cost stickiness. It should be noted that the variable $POST_t$ is omitted from the model due to perfect collinearity with the additional controls, since their values are zero before the outbreak of COVID-19. Furthermore, the results reveal that the coefficient on $\ln ES_n x \Delta \ln SALES_{i,t}$ is positive and statistically significant ($\delta_{4,g} = 0.133$, $p < 0.1$), indicating a positive relation between governmental economic support and cost elasticity for sales increases. This might be the case as managers expect economic support to enhance customers' purchasing power, which reduces their pessimism about future demand and thus results in increased resource extension, such as (re-)hirings, during periods of sales growth. All other additional control variables as well as their interaction terms, except for $\ln ES_n$ and $\ln ES_n xDEC_{i,t}$ which cannot be interpreted economically meaningful, lack statistical significance.

Sample split

After analysing the association between a formal institution and labour cost stickiness, this study investigates whether an informal institution, namely national culture, has a moderating role in this relation. If evidence in line with national culture moderating the relation between the stringency of government interventions during COVID-19 and labour cost stickiness is found, this could also explain why the results in *Table 6* show a positive association between stringent government interventions and cost stickiness, although theory provides contradictory predictions regarding the direction.

Model (2) is re-estimated for the high- and low-subsample for each dimension of national culture, i.e. PDI, IDV, MAS, UAI, LTO and IVR. *Table 8* presents the results. Similarly to the findings shown in *Table 4* and *Table 6*, the coefficients on $\Delta \ln SALES_{i,t}$ (δ_0) and $\Delta \ln SALES_{i,t} xDEC_{i,t}$ (θ_0) are significantly positive and negative, respectively, across both the high and the low subsample for most of the dimensions of national culture[38]. This reconfirms the prior finding that labour costs are on average sticky. However, the size of the coefficient on $\Delta \ln SALES_{i,t} xDEC_{i,t}$ (θ_0) differs between the high- and the low-subsample for some dimensions of national culture, indicating that national culture is directly related to the level of labour cost stickiness. These findings are in line with Banker et al. (2014, p. 224) who state that while behavioural factors, such as culture, "do not affect the general structure of asymmetric cost behaviour, they can either accentuate or diminish its magnitude". The coefficients diverge most in terms of absolute magnitude between the high- and low-subsample of the dimensions MAS ($\theta_{0,\,high} = -0.209$, $p < 0.05$ and $\theta_{0,\,low} = -0.366$, $p < 0.05$), UAI ($\theta_{0,\,high} = -0.230$, $p < 0.05$ and $\theta_{0,\,low} = -0.305$, $p < 0.05$) and LTO ($\theta_{0,\,high} = -0.072$, $p > 0.1$ and $\theta_{0,\,low} = -0.290$, $p < 0.01$). Thus, labour cost stickiness is found to be lower in firms headquartered in countries which exhibit high levels in masculinity, uncertainty-avoidance or long-term orientation. This aligns with the results of Kitching et al. (2016), who report a negative relation between cost stickiness and masculinity, high uncertainty avoidance, and long-term orientation as well. They explain the findings by stating that in masculine societies, managers tend to be "assertive, tough, and focused on material success" (Kitching et al., 2016, p. 406), which reduces psychological adjustment costs, such as those associated with firing employees, leading to lower levels of cost stickiness. Furthermore, decision-makers with high uncertainty avoidance are more loss-averse and prone to anticipating worst-case scenarios for future demand, resulting in decreased cost stickiness. Finally, Kitching et al. (2016) argue, that managers with a strong long-term orientation have a strong propensity for thriftiness (Hofstede et al., 2010) and are thus more inclined to cut costs when sales decline. However, it must be noted that the results of a z-Test, which assesses the statistical significance of the difference between the two coefficients $\theta_{0,\,high}$ and $\theta_{0,\,low}$, are only significant for the dimensions MAS and LTO and only if the regression is re-estimated without robust or clustered standard errors. *Table 8* displays the z-values and the associated statistical significance based on a two-tailed p-value both for the regression with robust and country-clustered standard errors and the regression without them (in brackets). The limited statistical significance of the differences may indicate a low economic significance, which must be considered when interpreting the results.

Next to the direct association between national culture and cost stickiness, the results suggests that national culture might moderate the relation between stringent government interventions during COVID-19 and labour cost stickiness, in contrast to Hypothesis 2 that claims the absence of a moderating relationship. The coefficient on the interaction term of interest $\Delta \ln SALES_{i,t} xDEC_{i,t} xTREAT_n xPOST_t$ (θ_3) varies between the high- and the low-subsample for most of the dimensions of national culture. For example, whereas θ_3 is statistically significantly negative for the low-subsample of PDI ($\theta_{3,\,high} = -0.165$, $p > 0.1$ and $\theta_{3,\,low} = -0.378$, $p < 0.01$), UAI ($\theta_{3,\,high} = -0.163$, $p > 0.1$ and $\theta_{3,\,low} = -0.381$, $p < 0.01$), LTO ($\theta_{3,\,high} = -0.122$, $p > 0.1$ and $\theta_{3,\,low} = -0.299$, $p < 0.1$) and IVR ($\theta_{3,\,high} = -0.238$, $p > 0.1$ and $\theta_{3,\,low} = -0.328$, $p < 0.05$), it has a smaller magnitude and is statistically insignificant for the high-subsamples of these cultural dimensions. Hence, the results indicate that the positive relation of stringent government interventions during COVID-19 and labour cost stickiness is amplified in firms headquartered in countries with low levels of power distance, uncertainty avoidance, long-term orientation or indulgence, but attenuated in firms headquartered in countries with high

[38] Only in the high-LTO-subsample θ_0 is not statistically significant.

levels of the respective dimensions. For the MAS dimension, the situation is reversed. Whereas the positive relation between stringent government interventions during COVID-19 and labour cost stickiness holds for cultures high in masculinity, it does not for more feminine ones ($\theta_{3,\,high} = -0.264$, $p < 0.05$ and $\theta_{3,\,low} = -0.715$, $p > 0.1$). Lastly, for the IDV dimension, the relation between stringent government interventions during COVID-19 and cost stickiness is negative and significant for both the high- and the low-subsample. However, the coefficient is smaller for the individualistic subsample than for the collectivistic one ($\theta_{3,\,high} = -0.283$ $p < 0.1$ and $\theta_{3,\,low} = -0.389$, $p < 0.05$).

The results largely align with the findings of Dheer et al. (2021) and J. Lee et al. (2024)[39]. Thus, they are also in line with the hypothesized argumentation that managers which accept stringent government interventions during COVID-19 as legitimate are more likely to have a positive outlook on future demand and are thus more inclined to keep resources in a situation of declining demand. In conclusion, managers' increased acceptance of governmental stringency results in increasing levels of cost stickiness. For the sake of brevity, not all dimensions of national culture are considered in detail, but an overview on the chain of argumentation is provided based on two dimensions[40]. For instance, as J. Lee et al. (2024) state, masculine cultures tend to prefer assertiveness and directive leadership and are therefore more likely to accept stringent policies as a legitimate response to the COVID-19 outbreak. Consequently, managers in masculine cultures might interpret the stringency of governmental interventions during COVID-19 as a positive signal which increases their optimism about future demand expectations. This is reflected by the results, which show that stringent government interventions during COVID-19 are positively related to labour cost stickiness only in more masculine, but not in more feminine countries. Furthermore, people in high uncertainty avoidance cultures are more averse towards change and risk, which makes them less willing to accept stringent government interventions (Dheer et al., 2021), as they constitute severe incisions in peoples' lives but are uncertain to be effective. Thus, managers in firms which are headquartered in countries high in uncertainty avoidance might question the inventions' legitimacy and thus draw less optimistic conclusions from stringent government interventions. Ultimately, this results in lower levels of labour cost stickiness, which is reflected in the coefficient of interests, θ_3, being insignificant in the high UAI subsample, but significant in the low UAI subsample.

Again, the z-Test which measures the between-sample difference in coefficients is only statistically significant in a regression specification without robust and clustered standard errors and only for the dimension MAS and UAI. This might raise doubts about the economic significance of the findings. Consequently, the results should not be interpreted causally, but only as an indication for a moderating role of national culture on the relation between stringent government interventions during COVID-19 and labour cost stickiness, especially considering the unequal sizes of the subsamples. Nevertheless, as the magnitude and the significance of the coefficients themselves is unchanged in both specifications, i.e. with and without robust and clustered standard errors, this study argues that overall, the results still support the notion that national culture moderates the relationship between stringent government interventions during COVID-19 and labour cost stickiness. Hence, Hypothesis 2, which claims the lack of a moderating role, is rejected.

Interestingly, these findings might also explain, as previously suggested, why the relationship between stringent government interventions during COVID-19 and labour cost stickiness is positive in the full sample, although theory predicts two opposing directions which were hypothesized to potentially cancel each other out. For example, the overall sample tends to be lower in PDI, UAI and LTO with less than 50% of the observations having a score above the median. However, this study's sample is comparably high in MAS, with 76 percent of the observations having a score above the median (which can be seen in *Table 2*). Hence, the full sample contains a comparably large number of observations from firms headquartered in countries having national cultures that are associated with a positive moderating role on the relationship between stringent government interventions and COVID-19, which could be the reason for the overall positive relation in the full sample[41].

5. Limitations and possibilities for future research

This study is subject to several limitations, most of them being related to the utilized DiD methodology. Firstly, the treatment variable is considered to be binary, despite being continuous in fact. Countries are assigned to either the treatment or the control group based on whether their average governmental stringency index during the COVID-19

[39] Whereas Dheer et al. (2021) examine the moderating effect of national culture on the relation between stringent government interventions and COVID-19 case growth, J. Lee et al. (2024) examine whether the six dimensions of culture strengthen or weaken the negative relation between stringent government interventions and mobility.

[40] It should be noted, that unlike for IDV, MAS, UAI and IVR, the results for the dimension PDI and LTO do not align with the expectations based on prior literature. Both Dheer et al. (2021) and J. Lee et al. (2024) argue that high power distance and long-term orientation is positively related to the acceptance of stringent government interventions. Opposingly, the findings of this study suggest that managers in low power distance countries and short-term oriented cultures derive more positive inferences from stringent government interventions, resulting in higher levels of labour cost stickiness. However, the results can be reconciled with Hofstede et al. (2010) who state that individuals in low power distance cultures are more inclined to pro-social acts like donating blood, and short-term oriented cultures put an emphasis on the fulfilment of social obligations.

[41] It should be noted that the sample also tends to be comparably high in IDV and IVR, with more than 50% of the observations having a score above the respective median. Both IDV and IVR are found to diminish the positive relation between the stringency of government interventions and labour cost stickiness. However, this is expected to be outweighed by the relatively large number of observations from low PDI, UAI and LTO and high MAS countries.

Table 8: Regression results Model (2) high- versus low-subsample of various dimensions of national culture

Variables	(I) PDI_high_n	(II) PDI_low_n	(III) IDV_high_n	(IV) IDV_low_n	(V) MAS_high_n	(VI) MAS_low_n	(VII) UAI_high_n	(VIII) UAI_low_n	(IX) LTO_high_n	(X) LTO_low_n	(XI) IVR_high_n	(XII) IVR_low_n
Main effects												
Constant ($ß_0$)	**0.036*****	**0.049*****	**0.050*****	0.011	**0.037*****	**0.064*****	**0.038*****	**0.047*****	**0.060*****	**0.034****	**0.045*****	**0.042****
	(0.009)	(0.014)	(0.010)	(0.015)	(0.010)	(0.013)	(0.009)	(0.014)	(0.007)	(0.013)	(0.012)	(0.016)
$TREAT_n$ ($ß_1$)	*omitted*	*omitted*	*omitted*	*omitted*	*omitted*	*omitted*	*omitted*	*omitted*	*omitted*	*omitted*	*omitted*	*omitted*
	(.)	*(.)*	*(.)*	*(.)*	*(.)*	*(.)*	*(.)*	*(.)*	*(.)*	*(.)*	*(.)*	*(.)*
$POST_t$ ($ß_2$)	-0.028	-0.032	-0.030	**-0.052****	**-0.039***	**-0.044***	-0.036	-0.031	-0.007	**-0.054***	-0.047	-0.024
	(0.032)	(0.024)	(0.025)	(0.022)	(0.019)	(0.022)	(0.026)	(0.025)	(0.014)	(0.028)	(0.027)	(0.023)
$SUC_DEC_{i,t}$ ($ß_{4,a}$)	-0.001	**-0.017****	-0.010	-0.012	**-0.017***	0.014	-0.005	**-0.016****	-0.003	-0.019	-0.013	-0.005
	(0.020)	(0.006)	(0.006)	(0.037)	(0.009)	(0.016)	(0.018)	(0.006)	(0.011)	(0.014)	(0.008)	(0.020)
$GDP_{n,t}$ ($ß_{4,b}$)	0.001	**-0.005*****	-0.003	-0.001	-0.003	0.001	0.001	**-0.005*****	0.002	**-0.006****	**-0.005****	0.001
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.004)	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.001)
$AINT_{i,t}$ ($ß_{4,c}$)	-0.015	0.021	**0.022***	-0.037	-0.004	**0.053***	-0.014	0.020	0.009	0.018	**0.029***	-0.018
	(0.023)	(0.014)	(0.012)	(0.034)	(0.013)	(0.027)	(0.021)	(0.015)	(0.011)	(0.020)	(0.016)	(0.021)
$EINT_{i,t}$ ($ß_{4,d}$)	**0.109*****	**0.071*****	**0.073*****	**0.120*****	**0.095*****	0.051	**0.111*****	**0.068*****	**0.087*****	**0.091*****	**0.075*****	**0.107*****
	(0.025)	(0.016)	(0.016)	(0.029)	(0.015)	(0.046)	(0.024)	(0.016)	(0.023)	(0.022)	(0.017)	(0.026)
$DEC_{i,t}$ (γ_0)	-0.027	**-0.020***	-0.018	**-0.029***	**-0.044*****	**-0.016***	-0.032	-0.018	-0.014	**-0.022****	-0.017	**-0.040****
	(0.023)	(0.010)	(0.011)	(0.013)	(0.011)	(0.008)	(0.022)	(0.010)	(0.023)	(0.008)	(0.010)	(0.016)
$\Delta \ln SALES_{i,t}$ (δ_0)	**0.403*****	**0.548*****	**0.566*****	**0.356*****	**0.389*****	**0.598*****	**0.400*****	**0.553*****	**0.424*****	**0.522*****	**0.549*****	**0.377*****
	(0.071)	(0.064)	(0.062)	(0.048)	(0.042)	(0.065)	(0.068)	(0.064)	(0.022)	(0.074)	(0.064)	(0.061)
Two-way interaction terms												
$TREAT_n x POST_t$ ($ß_3$)	0.006	-0.014	-0.014	0.011	0.010	0.012	0.004	-0.012	-0.017	0.013	-0.005	-0.001
	(0.021)	(0.017)	(0.012)	(0.027)	(0.017)	(0.029)	(0.021)	(0.018)	(0.014)	(0.015)	(0.011)	(0.022)
$TREAT_n x DEC_{i,t}$ (γ_1)	0.017	-0.007	-0.010	**0.066****	0.025	0.023	0.025	-0.009	0.004	-0.001	-0.016	**0.039***
	(0.030)	(0.013)	(0.014)	(0.027)	(0.017)	(0.015)	(0.029)	(0.013)	(0.023)	(0.018)	(0.013)	(0.018)
$TREAT_n x \Delta \ln SALES_{i,t}$ (δ_1)	0.132	**-0.242*****	**-0.195****	**0.253****	0.005	-0.067	0.137	**-0.251*****	0.009	-0.152	**-0.187****	0.073
	(0.092)	(0.077)	(0.086)	(0.091)	(0.071)	(0.125)	(0.090)	(0.077)	(0.091)	(0.107)	(0.078)	(0.122)

Table 8 — continued

$POST_t x DEC_{i,t}$ (γ_2)	**0.031**	**0.019**	**0.021***	**0.026**	**0.041*****	**0.006**	**0.031**	**0.019**	**0.008**	**0.028***	**0.021**	**0.042****
	(0.022)	**(0.012)**	**(0.011)**	**(0.022)**	**(0.013)**	**(0.013)**	**(0.023)**	**(0.012)**	**(0.020)**	**(0.013)**	**(0.013)**	**(0.015)**
$POST_t x \Delta \ln SALES_{i,t}$ (δ_2)	0.013	-0.061	-0.052	0.032	0.014	-0.079	0.009	-0.058	0.049	-0.056	-0.061	0.031
	(0.045)	(0.056)	(0.058)	(0.047)	(0.025)	(0.046)	(0.049)	(0.056)	(0.031)	(0.055)	(0.056)	(0.045)
$SUC_DEC_{i,t} x DEC_{i,t}$ ($\gamma_{4,a}$)	*omitted*	*omitted*	*omitted*	*omitted*	*omitted*	*omitted*	*omitted*	*omitted*	*omitted*	*omitted*	*omitted*	*omitted*
	(.)	*(.)*	*(.)*	*(.)*	*(.)*	*(.)*	*(.)*	*(.)*	*(.)*	*(.)*	*(.)*	*(.)*
$GDP_{n,t} x DEC_{i,t}$ ($\gamma_{4,b}$)	-0.002	**0.002****	0.001	**-0.003***	0.001	**-0.005***	-0.001	**0.002***	-0.000	0.001	0.001	-0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.002)
$AINT_{i,t} x DEC_{i,t}$ ($\gamma_{4,c}$)	0.013	0.007	**0.009****	0.007	**0.009***	0.011	0.009	**0.010***	0.005	0.009	0.008	0.013
	(0.010)	(0.006)	(0.004)	(0.019)	(0.005)	(0.009)	(0.010)	(0.005)	(0.007)	(0.007)	(0.005)	(0.010)
$EINT_{i,t} x DEC_{i,t}$ ($\gamma_{4,d}$)	0.010	0.002	0.005	0.006	0.011	-0.007	0.011	0.002	0.018***	-0.004	0.002	0.009
	(0.012)	(0.007)	(0.006)	(0.018)	(0.007)	(0.010)	(0.011)	(0.007)	(0.005)	(0.010)	(0.006)	(0.012)
$SUC_DEC_{i,t} x \Delta \ln SALES_{i,t}$ ($\delta_{4,a}$)	0.011	0.089	**0.092***	-0.026	0.020	**0.241*****	*omitted*	*omitted*	*omitted*	0.035	0.059	0.071
	(0.050)	(0.070)	(0.051)	(0.100)	(0.049)	(0.038)	*(.)*	*(.)*	*(.)*	(0.072)	(0.064)	(0.069)
$GDP_{n,t} x \Delta \ln SALES_{i,t}$ ($\delta_{4,b}$)	-0.003	**0.005****	0.002	-0.002	0.002	-0.009	-0.002	**0.004****	-0.003	0.002	0.002	0.001
	(0.004)	(0.002)	(0.003)	(0.004)	(0.003)	(0.009)	(0.004)	(0.002)	(0.006)	(0.003)	(0.004)	(0.003)
$AINT_{i,t} x \Delta \ln SALES_{i,t}$ ($\delta_{4,c}$)	**-0.108*****	**-0.092*****	**-0.111*****	-0.079	**-0.116*****	**-0.046****	**-0.109*****	**-0.087*****	**-0.147*****	**-0.071*****	**-0.097*****	**-0.109*****
	(0.032)	(0.014)	(0.017)	(0.045)	(0.017)	(0.019)	(0.031)	(0.015)	(0.017)	(0.016)	(0.015)	(0.034)
$EINT_{i,t} x \Delta \ln SALES_{i,t}$ ($\delta_{4,d}$)	**0.084*****	**0.077*****	**0.085*****	**0.052*****	**0.065*****	**0.131*****	**0.075****	**0.083*****	**0.094*****	**0.064****	**0.079*****	**0.079*****
	(0.027)	(0.020)	(0.021)	(0.014)	(0.018)	(0.017)	(0.026)	(0.021)	(0.023)	(0.022)	(0.024)	(0.021)
$\Delta \ln SALES_{i,t} x DEC_{i,t}$ (θ_0)	**-0.240****	**-0.292****	**-0.296****	**-0.276****	**-0.209****	**-0.366****	**-0.230****	**-0.305****	-0.072	**-0.290*****	**-0.264****	**-0.256****
	(0.092)	(0.099)	(0.102)	(0.112)	(0.074)	(0.124)	(0.092)	(0.099)	(0.176)	(0.087)	(0.102)	(0.094)
Three-way interaction terms												
$TREAT_n x POST_t x DEC_{i,t}$ (γ_3)	-0.016	-0.003	-0.004	-0.039	**-0.026***	-0.018	-0.017	-0.005	0.001	-0.019	-0.004	**-0.037***
	(0.026)	(0.014)	(0.012)	(0.043)	(0.012)	(0.047)	(0.026)	(0.015)	(0.022)	(0.016)	(0.015)	(0.019)
$TREAT_n x POST_t x \Delta \ln SALES_{i,t}$ (δ_3)	-0.037	**0.206****	**0.156***	-0.129	0.073	-0.061	-0.034	**0.205****	-0.002	0.152	0.147	0.027
	(0.058)	(0.066)	(0.075)	(0.116)	(0.052)	(0.221)	(0.059)	(0.066)	(0.069)	(0.085)	(0.084)	(0.078)
$TREAT_n x \Delta \ln SALES_{i,t} x DEC_{i,t}$ (δ_3)	0.012	**0.339****	**0.251***	**0.226***	0.155	**0.674***	0.011	**0.348****	-0.021	**0.258***	0.222	0.155
	(0.105)	(0.113)	(0.122)	(0.107)	(0.092)	(0.322)	(0.106)	(0.114)	(0.194)	(0.121)	(0.237)	(0.088)

Table 8 — continued

$POST_t x \Delta \ln SALES_{i,t} x DEC_{i,t}$ (θ_2)	**0.314*****	**0.265****	**0.268****	**0.255*****	**0.265*****	**0.230****	**0.308*****	**0.264****	0.151	**0.259****	**0.267****	**0.268****
	(0.069)	(0.097)	(0.104)	(0.071)	(0.030)	(0.095)	(0.061)	(0.097)	(0.098)	(0.094)	(0.101)	(0.102)
$SUC_DEC_{i,t} x \Delta \ln SALES_{i,t} x DEC_{i,t}$ ($\theta_{4,a}$)	*omitted*	*omitted*	*omitted*	*omitted*	*omitted*	*omitted*	0.002	0.096	**0.092****	*omitted*	*omitted*	*omitted*
	(.)	*(.)*	*(.)*	*(.)*	*(.)*	*(.)*	(0.050)	(0.071)	(0.035)	*(.)*	*(.)*	*(.)*
$GDP_{n,t} x \Delta \ln SALES_{i,t} x DEC_{i,t}$ ($\theta_{4,b}$)	-0.002	-0.003	-0.001	-0.014	-0.003	-0.015	-0.003	-0.002	-0.004	-0.003	-0.002	**-0.013***
	(0.006)	(0.003)	(0.003)	(0.011)	(0.003)	(0.015)	(0.006)	(0.003)	(0.007)	(0.003)	(0.003)	(0.007)
$AINT_{i,t} x \Delta \ln SALES_{i,t} x DEC_{i,t}$ ($\theta_{4,c}$)	0.035	-0.018	0.011	-0.034	0.017	-0.040	0.033	-0.018	0.033	-0.009	0.003	0.014
	(0.037)	(0.021)	(0.022)	(0.041)	(0.022)	(0.037)	(0.037)	(0.021)	(0.032)	(0.020)	(0.022)	(0.043)
$EINT_{i,t} x \Delta \ln SALES_{i,t} x DEC_{i,t}$ ($\theta_{4,d}$)	0.048	-0.028	-0.030	**0.172*****	0.030	-0.097	0.063	-0.042	0.012	-0.001	-0.014	0.029
	(0.038)	(0.042)	(0.038)	(0.032)	(0.027)	(0.054)	(0.037)	(0.045)	(0.038)	(0.052)	(0.044)	(0.048)
Four-way interaction term												
$\Delta \ln SALES_{i,t} x DEC_{i,t} x TREAT_n x POST_t$ (θ_3)	-0.165	**-0.378*****	**-0.283***	**-0.389****	**-0.264****	-0.715	-0.163	**-0.381*****	-0.122	**-0.299***	-0.238	**-0.328****
	(0.134)	(0.107)	(0.142)	(0.170)	(0.096)	(0.519)	(0.128)	(0.107)	(0.169)	(0.141)	(0.176)	(0.108)
Firm-fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year-fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Number of observations	5,864	9,582	12,618	2,828	11,791	3,655	6,144	9,302	7,365	8,081	9,774	5,672
R^2	0.268	0.305	0.311	0.264	0.266	0.378	0.264	0.310	0.266	0.296	0.308	0.250
Z-Value (difference of θ_0 between high and low subsample)	0.387		-0.131		1.085		0.551		1.116		-0.059	
	[0.521]		[-0.180]		**[1.642*****]**		[0.752]		**[1.750*****]**		[-0.079]	
Z-Value (difference of θ_3 between high and low subsample)	1.240		0.477		0.854		1.309		0.806		0.438	
	[1.630]		[0.593]		**[2.026******]**		**[1.686*****]**		[1.186]		[0.671]	

Robust standard errors clustered at the country-level are reported in parentheses.
*, **, *** indicate two-tailed significance at the 10 percent, 5 percent, and 1 percent levels.
Z-values in brackets are based on a regression specification without robust and clustered standard errors.

period is higher than the median of all sample countries' average stringency indices. This approach creates only two groups – the 'higher stringency group' and the 'lower stringency group' – although the actual variation in the degree of stringency is much broader. Secondly, it is assumed that the degree of stringency remains constant over the entire COVID-19 period. However, the stringency of government interventions may also change over time. For example, governments might impose restrictions cautiously at early stages when infections are low but increase stringency as COVID-19 cases rise (Pulejo & Querubín, 2021). To address these caveat, future research could utilize more sophisticated models, such as continuous (Morris et al., 2023) or staggered (Mader & Rüttenauer, 2022) DiD designs. Furthermore, a typical limitation of multinational samples is cross-country heterogeneity, which introduces an increased risk of omitted variable bias, thus making the comparison of firm-level measures across countries more difficult (Cannon et al., 2020; Ma et al., 2021). As shown in *Appendix 4*, covariate balance between treated and control firms prior to the treatment is questionable. The results of the t-test and the Kolmogorov-Smirnov test reveal statistically significant differences in the means and distributions of most variables, indicating that treated and control firms differ significantly in several observable characteristics during the pre-treatment period. Thus, the concern arises that firms headquartered in countries with more stringent government interventions might differ from firms headquartered in countries with less stringent government interventions in unobserved characteristics as well. Consequently, the study might be subject to omitted variables which could confound the results. Future studies might perform matching methods prior to conducting the analyses to improve covariate balance[42]. Additionally, although the study controls for pandemic affectedness and governmental economic support in a robustness check, there might be further time-varying confounding variables that influence treatment and control group differently. In summary, the results do not allow to draw conclusions about causal effects. As it is acknowledged that certain methodological issues remain unsolved, this study cannot go beyond demonstrating evidence on associations.

There are limitations unrelated to the DiD design as well. One caveat inherent in using financial data for cost stickiness research is that labour expenses and sales are imperfect proxies for economic labour costs and activity levels, because they might also be affected by accounting rules or price changes (Brüggen & Zehnder, 2014). Moreover, the study assumes that only the stringency of interventions in the firm's headquarters country is relevant for managers' decision-making. Though, many firms conduct business Europe- or worldwide and rely heavily on geographically expansive supply chains. Therefore, the stringency of government interventions in other countries could also influence managerial decision-making and, consequently, labour cost stickiness. To mitigate this issue, future research could apply this study's methodology to a sample of private instead of publicly listed firms, since they might operate more nationally. Furthermore, national culture may not necessarily influence managerial behaviour and decision-making in the assumed way. Neither is it clear that the managers' nationality or country of residence coincides with the location of the firm's headquarters, nor is it possible to rule out that managers' behaviour is also influenced by other factors such as corporate culture, education or past experiences (Kitching et al., 2016). As BenYoussef et al. (2023) and Brüggen and Zehnder (2014) suggest, future research could aim at assessing cognitive processes underlying managerial decisions more directly by using methods such as surveys or laboratory experiments. Lastly, this study is only able to present an indication of the moderating role of culture on the relation between stringent government interventions during COVID-19 and labour cost stickiness, as the statistical significance of the differences between the cultural characteristics is not given in all regression specifications. Thus, the results should be treated cautiously and motivate future research to deepen research in this field, for example by investigating whether the positive association between stringent government interventions and labour cost stickiness also exists for other regions with different cultural backgrounds, such as America or Asia.

6. Conclusion

This study is conducted to examine how formal and informal country-level institutions are associated with corporate cost behaviour within the COVID-19 context. Specifically, it aims at answering the research questions of how the stringency of government interventions during COVID-19 is related to labour cost stickiness and whether national culture moderates this relation. To accomplish that, the study investigates a sample of 15,446 firm-year observations of 3,383 publicly listed firms from 25 European countries in the period 2017 to 2022. First, this study empirically tests for the existence of labour cost stickiness in the sample and finds, in line with prior literature, that labour costs are sticky on average. Next, the level of labour cost stickiness before and after the outbreak of the COVID-19 pandemic in 2020 is compared. This study finds that the sticky behaviour of labour costs disappears after 2020, aligning with theory as well as previous research findings which state that the COVID-19-induced uncertainty increases managerial pessimism about future demand. Furthermore, this study provides evidence of a positive association between stringent government interventions during COVID-19 and labour cost stickiness, while controlling for known economic determinants of cost asymmetry. Using a DiD model, the study shows that firms headquartered in countries with stringent government interventions during the pandemic exhibit higher labour cost sensitivity towards

[42] This study refrains from employing matching methods due data availability constraints. As the sample is unbalanced and the sample period covers only six years, variance in the data is not large enough to compute a firm-level score for labour cost stickiness using Model (1) or Model (3) for a sufficiently high number of observations on which matching based on a labour cost stickiness score could be performed.

sales increases and lower labour cost sensitivity towards sales decreases after 2020, compared to firms headquartered in countries with less stringent government interventions during COVID-19. This is consistent with the argument that stringent government interventions signal the governments' ability to effectively manage the pandemic, which in turn increases managerial optimism about future demand conditions. Hence, managers from firms located in countries with stringent government interventions during COVID-19 are more likely to view sales growths as permanent and sales drops as temporary, resulting in increased levels of labour cost elasticity for sales increases and increased levels of labour cost stickiness for sales decreases. Finally, a sample split based on the median value of each of Hofstede's six dimensions of national culture, namely PDI, IDV, MAS, UAI, LTO and IVR, is conducted to examine whether culture moderates the relation between the stringency of government interventions during COVID-19 and labour cost stickiness. The results provide an initial indication for national culture indeed strengthening or weakening the association. This is consistent with the idea that depending on national culture, managers are more or less inclined to accept stringent government interventions, which in turn shapes their optimism about the future and consequently results in increased or decreased levels of labour cost stickiness. As such, the results of this study deepen the understanding of how managers make resource adjustment decisions in times of crises and why the level of labour cost stickiness varies across countries.

In addition to extending the academic literature on institutions and cost asymmetry, this study has practical implications as well. It provides governments with insights on how state-level interventions can influence managers' perceptions of the future, ultimately leading to operational firm-level consequences such as changes in labour resource management. Therefore, the results can offer guidance for governments when designing policy responses during potential future pandemics, allowing them to consider economic, next to health-related consequences. By leveraging the positive relationship between the stringency of interventions and the exhibited levels of corporate labour cost stickiness, governments might be able to mitigate costs related to layoffs and possible resulting social unrest to some extent (Gu et al., 2020). This supplies them with additional arguments to justify stringent interventions to businesses and citizens. The results of this study also show that governments should consider national culture when planning actions to combat the spread of viruses, as the cultural context might influence whether and to what degree stringent governmental interventions are able to shape managers' future expectations and, consequently, resource management decisions. Thus, accounting for cultural diversity by allowing variance in the degree of stringency is particularly crucial when developing transnational policies, such as EU-wide regulations.

Although causal relationships cannot be detected in this study's setting, it still provides valuable insights that can serve as a foundation for further academic discussions and practical considerations on firm-level cost implications of stringent government interventions during pandemics in distinct cultural settings.

References

Aggarwal, S., Nawn, S., & Dugar, A. (2021). What caused global stock market melt-down during the COVID pandemic–Lockdown stringency or investor panic? *Finance Research Letters*, *38*, Article 101827. https://doi.org/10.1016/j.frl.2020.101827

Akter, S. (2020). The impact of COVID-19 related 'stay-at-home' restrictions on food prices in Europe: findings from a preliminary analysis. *Food Security*, *12*(4), 719–725. https://doi.org/10.1007/s12571-020-01082-3

Alfano, V., Ercolano, S., & Pinto, M. (2022). Carrot and stick: Economic support and stringency policies in response to COVID-19. *Evaluation and Program Planning*, *94*, Article 102129. https://doi.org/10.1016/j.evalprogplan.2022.102129

Altig, D., Baker, S., Barrero, J. M., Bloom, N., Bunn, P., Chen, S., Davis, S. J., Leather, J., Meyer, B., Mihaylov, E., Mizen, P., Parker, N., Renault, T., Smietanka, P., & Thwaites, G. (2020). Economic uncertainty before and during the COVID-19 pandemic. *Journal of Public Economics*, *191*, Article 104274. https://doi.org/10.1016/j.jpubeco.2020.104274

Anderson, M. C., Banker, R. D., & Janakiraman, S. N. (2003). Are Selling, General, and Administrative Costs "Sticky"? *Journal of Accounting Research*, *41*(1), 47–63. https://doi.org/10.1111/1475-679X.00095

Anderson, S. W., & Lanen, W. N. (2007). Understanding Cost Management: What Can We Learn from the Evidence on 'Sticky Costs'? https://ssrn.com/abstract=975135

Ando, S., Balakrishnan, R., Gruss, B., Hallaert, J.-J., Jirasavetakul, L.-B. F., Kirabaeva, K., Klein, N., Lariau, A., Liu, L. Q., Malacrino, D., Qu, H., & Solovyeva, A. (2022). *European Labor Markets and the COVID-19 Pandemic: Fallout and the Path Ahead (Departmental Paper No. 2022/004)* (tech. rep.). International Monetary Fund. https://www.imf.org/en/Publications/Departmental-Papers-Policy-Papers/Issues/2022/03/02/European-Labor-Markets-and-the-COVID-19-Pandemic-Fallout-and-the-Path-Ahead-512327

Ashraf, B. N. (2020). Economic impact of government interventions during the COVID-19 pandemic: International evidence from financial markets. *Journal of Behavioral and Experimental Finance*, *27*, Article 100371. https://doi.org/10.1016/j.jbef.2020.100371

Atanasov, V. A., & Black, B. S. (2016). Shock-based causal inference in corporate finance and accounting research. *Critical Finance Review*, *5*(2), 207–304. https://doi.org/10.1561/104.00000036

Bakry, W., Kavalmthara, P. J., Saverimuttu, V., Liu, Y., & Cyril, S. (2022). Response of stock market volatility to COVID-19 announcements and stringency measures: A comparison of developed and emerging markets. *Finance Research Letters*, *46*(Part A), Article 102350. https://doi.org/10.1016/j.frl.2021.102350

Balakrishnan, R., & Gruca, T. (2008). Cost stickiness and core competency: A note. *Contemporary Accounting Research*, *25*(4), 993–1006. https://doi.org/10.1506/car.25.4.2

Banker, R. D., & Byzalov, D. (2014). Asymmetric Cost Behavior. *Journal of Management Accounting Research*, *26*(2), 43–79. https://doi.org/10.2308/jmar-50846

Banker, R. D., Byzalov, D., & Chen, L. T. (2013). Employment protection legislation, adjustment costs and cross-country differences in cost behavior. *Journal of Accounting and Economics*, *55*(1), 111–127. https://doi.org/10.1016/j.jacceco.2012.08.003

Banker, R. D., Byzalov, D., Ciftci, M., & Mashruwala, R. (2014). The Moderating Effect of Prior Sales Changes on Asymmetric Cost Behavior. *Journal of Management Accounting Research*, *26*(2), 221–242. https://doi.org/10.2308/jmar-50726

Banker, R. D., Byzalov, D., Fang, S., & Liang, Y. (2018). Cost Management Research. *Journal of Management Accounting Research*, *30*(3), 187–209. https://doi.org/10.2308/jmar-51965

Banker, R. D., Fang, S., & Mehta, M. N. (2020). Anomalous Operating Performance During Economic Slowdowns. *Journal of Management Accounting Research*, *32*(2), 57–83. https://doi.org/10.2308/jmar-52547

Belina, H., Surysekar, K., & Weismann, M. (2019). On the medical loss ratio (MLR) and sticky selling general and administrative costs: Evidence from health insurers. *Journal of Accounting and Public Policy*, *38*(1), 53–61. https://doi.org/10.1016/j.jaccpubpol.2019.01.004

BenYoussef, N., Bouslimi, L., & Khan, S. (2023). *Navigating Cost Behaviour during COVID-19 Uncertainty: Financial Flexibility, Managerial Traits, and Government Interventions* (tech. rep.). SSRN Working Paper Nr. 4633434. https://ssrn.com/abstract=4633434

Biddlestone, M., Green, R., & Douglas, K. M. (2020). Cultural orientation, power, belief in conspiracy theories, and intentions to reduce the spread of COVID-19. *British Journal of Social Psychology*, *59*(3), 663–673. https://doi.org/10.1111/bjso.12397

Brüggen, A., & Zehnder, J. O. (2014). SG&A cost stickiness and equity-based executive compensation: does empire building matter? *Journal of Management Control*, *25*, 169–192. https://doi.org/10.1007/s00187-014-0195-5

Buchheim, L., Dovern, J., Krolage, C., & Link, S. (2022). Sentiment and firm behavior during the COVID-19 pandemic. *Journal of Economic Behavior and Organization*, *195*, 186–198. https://doi.org/10.1016/j.jebo.2022.01.011

Bugeja, M., Lu, M., & Shan, Y. (2015). Cost Stickiness in Australia: Characteristics and Determinants. *Australian Accounting Review*, *25*(3), 248–261. https://doi.org/10.1111/auar.12066

Caggiano, G., Castelnuovo, E., & Kima, R. (2020). The global effects of Covid-19-induced uncertainty. *Economic Letters*, *194*, Article 109392. https://doi.org/10.1016/j.econlet.2020.109392

Calleja, K., Steliaros, M., & Thomas, D. C. (2006). A note on cost stickiness: Some international comparisons. *Management Accounting Research*, *17*(2), 127–140. https://doi.org/10.1016/j.mar.2006.02.001

Cannon, J. N., Hu, B., Lee, J. J., & Yang, D. (2020). The effect of international takeover laws on corporate resource adjustment: Market discipline and/or managerial myopia? *Journal of International Business Studies*, *51*, 1443–1477. https://doi.org/10.1057/s41267-020-00370-6

Chen, C. X., Lu, H., & Sougiannis, T. (2012). The Agency Problem, Corporate Governance, and the Asymmetrical Behavior of Selling, General, and Administrative Costs. *Contemporary Accounting Research*, *29*(1), 252–282. https://doi.org/10.1111/j.1911-3846.2011.01094.x

Chen, C. X., Nasev, J., & Wu, S. Y.-C. (2022). CFO Overconfidence and Cost Behavior. *Journal of Management Accounting Research*, *34*(2), 117–135. https://doi.org/10.2308/JMAR-18-055

Chen, L., Huang, J., Hu, D., & Chen, X. (2022). Dividend regulation and cost stickiness: evidence from a quasi-natural experiment. *China Accounting and Finance Review*, *24*(4), 486–515. https://doi.org/10.1108/CAFR-08-2022-0092

Ciftci, M., & Zoubi, T. A. (2019). The Magnitude of Sales Change and Asymmetric Cost Behavior. *Journal of Management Accounting Research*, *31*(3), 65–81. https://doi.org/10.2308/jmar-52331

Clogg, C. C., Petkova, E., & Haritou, A. (1995). Statistical Methods for Comparing Regression Coefficients Between Models. *American Journal of Sociology*, *100*(5), 1261–1293. https://doi.org/10.1086/230638

Cooper, R., & Kaplan, R. S. (1992). Activity-Based Systems: Measuring the Costs of Resource Usage. *Accounting Horizons*, *6*(3), 1–13.

Dalla Via, N., & Perego, P. (2014). Sticky cost behaviour: evidence from small and medium-sized companies. *Accounting & Finance*, *54*(3), 753–778. https://doi.org/10.1111/acfi.12020

Dheer, R. J. S., Egri, C. P., & Treviño, L. J. (2021). A cross-cultural exploratory analysis of pandemic growth: The case of COVID-19. *Journal of International Business Studies*, *52*, 1871–1892. https://doi.org/10.1057/s41267-021-00455-w

Dierynck, B., Landsman, W. R., & Renders, A. (2012). Do Managerial Incentives Drive Cost Behavior? Evidence about the Role of the Zero Earnings Benchmark for Labor Cost Behavior in Private Belgian Firms. *The Accounting Review*, *87*(4), 1219–1246. https://doi.org/10.2308/accr-50153

Ekici, E., Zhang, J., & Han, J. (2024). The impact of cost resilience on management decisions to respond to the COVID-19 crisis in healthcare companies. *International Journal of Healthcare Management*, *1-9*. https://doi.org/10.1080/20479700.2024.2318512

Feng, G.-F., Yang, H.-C., Gong, Q., & Chang, C.-P. (2021). What is the exchange rate volatility response to COVID-19 and government interventions? *Economic Analysis and Policy*, *69*, 705–719. https://doi.org/10.1016/j.eap.2021.01.018

Finne, G., Haga, J., & Sundvik, D. (2024). Cost behavior and reporting frequency during the COVID-19 outbreak. *Accounting and Business Research*, *54*(4), 491–520. https://doi.org/10.1080/00014788.2023.2298784

Ghazy, W., Rahman, A., Kabir, H., & Scott, T. (2024). *Cost Stickiness and Covid-19 Extreme Uncertainty* (tech. rep.). SSRN Working Paper Nr. 4777381. https://ssrn.com/abstract=4777381

Goodman-Bacon, A., & Marcus, J. (2020). Using Difference-in-Differences to Identify Causal Effects of COVID-19 Policies. *Survey Research Methods*, *14*(2), 153–158. https://doi.org/10.18148/srm/2020.v14i2.7723

Graafland, J., & Noorderhaven, N. (2020). Culture and institutions: How economic freedom and long-term orientation interactively influence corporate social responsibility. *Journal of International Business Studies*, *51*, 1034–1043. https://doi.org/10.1057/s41267-019-00301-0

Gu, Z., Tang, S., & Wu, D. (2020). The Political Economy of Labor Employment Decisions: Evidence from China. *Management Science*, *66*(10), 4703–4725. https://doi.org/10.1287/mnsc.2019.3345

Hale, T., Angrist, N., Goldszmidt, R., Kira, B., Petherick, A., Phillips, T., Webster, S., Cameron-Blake, E., Hallas, L., Majumdar, S., & Tatlow, H. (2021). A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). *Nature Human Behaviour*, *5*, 529–538. https://doi.org/10.1038/s41562-021-01079-8

Hartlieb, S., Loy, T., & Eierle, B. (2020a). Does community social capital affect asymmetric cost behaviour? *Management Accounting Research*, *46*, Article 100640. https://doi.org/10.1016/j.mar.2019.02.002

Hartlieb, S., Loy, T., & Eierle, B. (2020b). The effect of generalized trust on cost stickiness: Cross-country evidence. *The International Journal of Accounting*, *55*(4), Article 2050018. https://doi.org/10.1142/S1094406020500183

Hassanein, A., & Younis, M. (2020). Cost Stickiness Behavior and Financial Crisis: Evidence from the UK Chemical Industry. *Corporate Ownership and Control*, *17*(2), 46–56. https://doi.org/10.22495/cocv17i2art4

Hausman, J. A. (1978). Specification Tests in Econometrics. *Econometrica*, *46*(6), 1251–1271. https://www.jstor.org/stable/1913827

Helmke, G., & Levitsky, S. (Eds.). (2006). *Informal Institutions and Democracy: Lessons from Latin America*. Johns Hopkins University Press. https://doi.org/10.56021/9780801883514

Hodgson, G. M. (2006). What Are Institutions? *Journal of Economic Issues*, *40*(1), 1–25. https://doi.org/10.1080/00213624.2006.11506879

Hofstede, G. H., Hofstede, G. J., & Minkov, M. (2010). *Cultures and Organizations, Software of the Mind: Intercultural Cooperation and Its Importance for Survival*. McGraw-Hill Professional.

Holzhacker, M., Krishnan, R., & Mahlendorf, M. D. (2015a). The Impact of Changes in Regulation on Cost Behavior. *Contemporary Accounting Research*, *32*(2), 534–566. https://doi.org/10.1111/1911-3846.12082

Holzhacker, M., Krishnan, R., & Mahlendorf, M. D. (2015b). Unraveling the Black Box of Cost Behavior: An Empirical Investigation of Risk Drivers, Managerial Resource Procurement, and Cost Elasticity. *The Accounting Review*, *90*(6), 2305–2335. https://doi.org/10.2308/accr-51092

Ibrahim, A. E. A. (2015). Economic Growth and Cost Stickiness: Evidence from Egypt. *Journal of Financial Reporting and Accounting*, *13*(1), 119–140. https://doi.org/10.1108/JFRA-06-2014-0052

Ibrahim, A. E. A., Ali, H., & Aboelkheir, H. (2022). Cost Stickiness: A Systematic Literature Review of 27 Years of Research and a Future Research Agenda. *Journal of International Accounting, Auditing and Taxation*, *46*, Article 100439. https://doi.org/10.1016/j.intaccaudtax.2021.100439

Inglehart, R., & Oyserman, D. (2004). Individualism, Autonomy, and Self-Expression: The Human Development Syndrome. In H. Vinken, J. Soeters, & P. Ester (Eds.), *Comparing Cultures: Dimensions of Culture in a Comparative Perspective* (pp. 74–96). Brill.

Jensen, M. C., & Meckling, W. H. (1976). Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure. *Journal of Financial Economics*, *3*(4), 305–360. https://doi.org/10.1016/0304-405X(76)90026-X

Jiménez, J., & Perdiguero, J. (2019). Difference-in-Difference. In A. Marciano & G. B. Ramello (Eds.), *Encyclopedia of Law and Economics* (pp. 551–555). Springer. https://doi.org/10.1007/978-1-4614-7753-2_664

Jin, X., & Wu, H. (2021). Economic Policy Uncertainty and Cost Stickiness. *Management Accounting Research*, *52*, Article 100750. https://doi.org/10.1016/j.mar.2021.100750

Kama, I., & Weiss, D. (2013). Do Earnings Targets and Managerial Incentives Affect Sticky Costs? *Journal of Accounting Research*, *51*(1), 201–224. https://doi.org/10.1111/j.1475-679X.2012.00471.x

Karatzimas, S., Naoum, V. C., & Narbón-Perpiña, I. (2022). Municipal Cost Behavior during Financial Crisis and Financial Sustainability Periods: Evidence from Spain. *International Public Management Journal*, *25*(6), 862–882. https://doi.org/10.1080/10967494.2021.1947924

Kitching, K., Mashruwala, R., & Pevzner, M. (2016). Culture and Cost Stickiness: A Cross-country Study. *The International Journal of Accounting*, *51*(3), 402–417. https://doi.org/10.1016/j.intacc.2016.07.010

Kizys, R., Tzouvanas, P., & Donadelli, M. (2021). From COVID-19 Herd Immunity to Investor Herding in International Stock Markets: The Role of Government and Regulatory Restrictions. *International Review of Financial Analysis*, *74*, Article 101663. https://doi.org/10.1016/j.irfa.2021.101663

Kuo, N. T., & Lee, C. F. (2023). Political Institutions and Cost Stickiness: International Evidence. *European Accounting Review*, *32*(3), 745–778. https://doi.org/10.1080/09638180.2021.2000459

Kwak, J.-Y., Lee, J.-W., & Jo, E. H. (2021). Impact of COVID-19 on R&D Cost Stickiness in IT Industry. *International Journal of Internet, Broadcasting and Communication*, *13*(2), 36–42. https://doi.org/10.7236/IJIBC.2021.13.2.36

Lee, J., Kim, O., & Park, C. (2024). A Cross-Cultural Study on the Effects of Government Control Policies on Mobility in COVID-19. *Cross-Cultural Research*, *0*(0), 1–28. https://doi.org/10.1177/10693971241245878

Lee, W. J., Pittman, J., & Saffar, W. (2020). Political Uncertainty and Cost Stickiness: Evidence from National Elections around the World. *Contemporary Accounting Research*, *37*(2), 1107–1139. https://doi.org/10.1111/1911-3846.12547

Liu, Y., & Jin, M. (2023). Does Supply Chain Network Centrality Affect Firm Cost Stickiness? *Finance Research Letters*, *58*(Part B), Article 104459. https://doi.org/10.1016/j.frl.2023.104459

Ma, L., Wang, X., & Zhang, C. (2021). Does Religion Shape Corporate Cost Behavior? *Journal of Business Ethics*, *170*, 835–855. https://doi.org/10.1007/s10551-019-04377-4

Mader, S., & Rüttenauer, T. (2022). The Effects of Non-pharmaceutical Interventions on COVID-19 Mortality: A Generalized Synthetic Control Approach Across 169 Countries. *Frontiers in Public Health*, *10*, Article 820642. https://doi.org/10.3389/fpubh.2022.820642

Meyer, B. H., Prescott, B., & Sheng, X. S. (2022). The Impact of the COVID-19 Pandemic on Business Expectations. *International Journal of Forecasting*, *38*(2), 529–544. https://doi.org/10.1016/j.ijforecast.2021.02.009

Morris, W., Correa, A., & Leiva, R. (2023). Impact of COVID-19 Containment Measures on Unemployment: A Multi-country Analysis Using a Difference-in-Differences Framework. *International Journal of Health Policy and Management*, *12*(1), 1–9. https://doi.org/10.34172/ijhpm.2022.7036

Noreen, E. (1991). Conditions under which Activity-based Cost Systems Provide Relevant Costs. *Journal of Management Accounting Research*, *3*, 159–168.

Noreen, E., & Soderstrom, N. (1994). Are Overhead Costs Strictly Proportional to Activity? Evidence from Hospital Departments. *Journal of Accounting and Economics*, *17*(1-2), 255–278. https://doi.org/10.1016/0165-4101(94)90012-4

Noreen, E., & Soderstrom, N. (1997). The Accuracy of Proportional Cost Models: Evidence from Hospital Service Departments. *Review of Accounting Studies*, *2*, 89–114. https://doi.org/10.1023/A:1018325711417

North, D. C. (1990). *Institutions, Institutional Change and Economic Performance*. Cambridge University Press. https://doi.org/10.1017/CBO9780511808678

Pinnuck, M., & Lillis, A. M. (2007). Profits versus Losses: Does Reporting an Accounting Loss Act as a Heuristic Trigger to Exercise the Abandonment Option and Divest Employees? *The Accounting Review*, *82*(4), 1031–1053. https://doi.org/10.2308/accr.2007.82.4.1031

Prabowo, R., Hooghiemstra, R., & Van Veen-Dirks, P. (2018). State Ownership, Socio-political Factors, and Labor Cost Stickiness. *European Accounting Review*, *27*(4), 771–796. https://doi.org/10.1080/09638180.2017.1329659

Pulejo, M., & Querubín, P. (2021). Electoral Concerns Reduce Restrictive Measures during the COVID-19 Pandemic. *Journal of Public Economics*, *198*, Article 104387. https://doi.org/10.1016/j.jpubeco.2021.104387

Redmond, W. H. (2005). A Framework for the Analysis of Stability and Change in Formal Institutions. *Journal of Economic Issues*, *39*(4), 665–681. https://doi.org/10.1080/00213624.2005.11506839

Saif-Alyousfi, A. Y. H. (2022). The Impact of COVID-19 and the Stringency of Government Policy Responses on Stock Market Returns Worldwide. *Journal of Chinese Economic and Foreign Trade Studies*, *15*(1), 87–105. https://doi.org/10.1108/JCEFTS-07-2021-0030

Stimolo, M. I., & Porporato, M. (2019). How Different Cost Behaviour Is in Emerging Economies? Evidence from Argentina. *Journal of Accounting in Emerging Economies*, *10*(1), 21–47. https://doi.org/10.1108/JAEE-05-2018-0050

Subramaniam, C., & Weidenmier Watson, M. (2016). Additional Evidence on the Sticky Behavior of Costs. *Advances in Management Accounting*, *26*, 275–305. https://doi.org/10.1108/S1474-787120150000026006

Sunstein, C. R. (1996). Social Norms and Social Roles. *Columbia Law Review*, *96*(4), 903–968. https://doi.org/10.2307/1123430

Wang, Z., Li, Y., Xu, R., & Yang, H. (2022). How Culture Orientation Influences the COVID-19 Pandemic: An Empirical Analysis. *Frontiers in Psychology*, *13*, 1–15. https://doi.org/10.3389/fpsyg.2022.899730

Weiss, D. (2010). Cost Behavior and Analysts' Earnings Forecasts. *The Accounting Review*, *85*(4), 1441–1471. https://doi.org/10.2308/accr.2010.85.4.1441

World Health Organization. (2020). WHO Director-General's Opening Remarks at the Media Briefing on COVID-19 - 11 March 2020. https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020

Yang, D. (2015). Mergers, CEO Hubris, and Cost Stickiness. *Emerging Markets Finance and Trade*, *51*(5), 46–63. https://doi.org/10.1080/1540496X.2015.1062313

Yang, H., & Deng, P. (2021). The Impact of COVID-19 and Government Interventions on Stock Markets of OECD Countries. *Asian Economic Letters*, *1*(4), 1–6. https://doi.org/10.46557/001c.18646

Yang, Y., & Chen, D. (2024). Influence of COVID-19 on Asymmetric Cost Behavior and Intellectual Capital Efficiency: A Comparison of Australian and Chinese Listed Firms. *Asia-Pacific Journal of Accounting & Economics*, *31*(3), 477–493. https://doi.org/10.1080/16081625.2023.2194887

Zaremba, A., Kizys, R., Aharon, D. Y., & Demir, E. (2020). Infected Markets: Novel Coronavirus, Government Interventions, and Stock Return Volatility around the Globe. *Finance Research Letters*, *35*, Article 101597. https://doi.org/10.1016/j.frl.2020.101597

Zhang, R., Hora, M., John, S., & Wier, H. A. (2022). Competition and Slack: The Role of Tariffs on Cost Stickiness. *Journal of Operations Management*, *68*(8), 855–880. https://doi.org/10.1002/joom.1162

Junior Management Science 10(3) (2025) 748-780



Junior Management Science

www.jums.academy
ISSN: 2942-1861



Modeling the Impact of Emission Credit Systems on Automotive Product Portfolios: A Mathematical Analysis of Policy Effects in Europe, China, and the U.S. Under Different Demand Scenarios

Zewei Shi

Technical University of Munich

Abstract

In the midst of the global climate crisis, governments worldwide have implemented a range of emission policies aimed at encouraging more production of the environmentally friendly vehicle. However, the exact impact of these policies on automakers' production portfolios and profitability remains uncertain and challenging to anticipate. This paper presents a comprehensive analysis of three major emission regulation policies enacted by the European Union (EU), China, and the United States (U.S.), evaluating their influence on car manufacturers. Leveraging a mathematical model, this paper adopt the perspective of individual manufacturers seeking to maximize revenue, delving into the intricacies of these policies. Furthermore, this article conduct sensitivity and factorial analyses to assess the impact of policy parameters. The findings reveal that all three major emission policies contribute to an increase in the production of low-emission vehicles. However, China's policy has the least impact on manufacturers' profits and relies more on market demand to reduce the average carbon fleet emissions compared to the policies in the EU and the U.S. In conclusion, this paper underscores that different policy systems yield varying profit outcomes for manufacturers, necessitating adjustments to production portfolios for sustained profitability and the significance of mathematical models in aiding manufacturers' understanding of evolving policies and making informed predictions in a dynamic regulatory landscape.

Keywords: automotive production; green transition; international emission policies; regulatory impact; sustainability

1. Introduction

As modern industrialization surges forward, humanity confronts the complex challenges of climate change. This encompasses the onset of extreme weather patterns and elevated temperatures, both driven by the incessant release of copious amounts of greenhouse gases into the atmosphere. The excessive emissions of greenhouse gases, such as carbon dioxide (CO_2) and methane (CH_4), instigate the greenhouse effect, culminating in the far-reaching issue of global warming. This phenomenon poses a threat to the existing ecosystem, manifesting in disruptive weather patterns and extreme climatic events (Yoro & Al., 2020). Notably, the primary source of CO_2 emissions stems not only from industrial production but also from vehicular exhaust (Huang et al., 2015). Traditional vehicles predominantly powered by gasoline and diesel generate substantial CO_2 emissions in day-to-day usage. In response to this environmental challenge, the electric vehicle concept emerges as a viable solution. By utilizing electricity as the primary power source, electric vehicles could reduce carbon emissions, positioning them as a more eco-friendly alternative (Costa et al., 2021)

Currently, there are four main types of vehicles: Internal Combustion Engine Vehicles (ICEV), Plug-in Hybrid Vehicles (PHEV), Battery Electric Vehicles (BEV), and Fuel Cell Elec-

I would like to sincerely thank my supervisor, Maximilian Kolter for his thoughtful guidance and support throughout my thesis. He not only helped me approach the problem more effectively but also provided valuable feedback to improve the clarity and readability of my writing. I am also grateful to Professor Kolish for providing the data and initial literature that laid the groundwork for my research.

DOI: https://doi.org/10.5282/jums/v10i3pp748-780

© The Author(s) 2025. Published by Junior Management Science.
This is an Open Access article distributed under the terms of the CC-BY-4.0 (Attribution 4.0 International). Open Access funding provided by ZBW.

tric Vehicles (ICEV) (Vacheva & Hinov, 2019). ICEV cars are the most traditional, powered by gasoline and emitting a relatively high amount of CO2. BEV and FCEV are pure electric vehicles with zero carbon emissions, collectively referred to as ZEV (Zero Emission Vehicles). The distinction between these two vehicle types lies in their power sources: BEVs use batteries charged with electricity, while FCEVs utilize biofuels and hydrogen-powered fuel cells, which can also be considered hydrogen-powered vehicles (Parikh et al., 2023). PHEVs represent a middle ground, with lower carbon emissions compared to ICEVs. They operate on a hybrid power source, utilizing both electricity and petrol or gasoline simultaneously. These different vehicle types have significant variations in production costs and sales revenue, depending on factors such as production year, type, and size (Lipman & Delucchi, 2006).

The ZEV and PHEV vehicle types produce less CO2 with lower tailpipe emissions, which can help mitigate the problem of global warming. However, despite the increasing popularity of electric and low-carbon emission vehicles, the cost of new energy vehicles remains higher than traditional vehicles, leading to lower profits for manufacturers (Cuenca et al., 2000). Vehicle producers need to maintain market competitiveness and prioritize profit and financial gains, making it challenging to persuade them to prioritize the production of less profitable but environmentally friendly vehicles. To encourage manufacturers to reduce fleet emissions, which represent the average amount of CO2 produced across their production portfolio, various governments have introduced policies and regulations. Different governments have adopted unique approaches to establish country-specific policies (An et al., 2011). However, the precise effectiveness of these policy systems on manufacturers' production portfolios remains unclear.

This article focuses on three primary markets: Europe, China, and the USA, each with its own emission regulations. In Europe, the emission policy is referred to as the Super Credit Policy, in China, it is known as the Dual Credit Policy, and in the USA, it is named the US Credit Policy. Further details about these regulations can be found in Section 3.4.

In this scenario, two stakeholders exist: the manufacturer and the policymaker. For the manufacturer, they need to determine quickly how different policies will affect their business profit and make quick responses to the production portfolio to ensure profitability. For the policymaker, they need to balance the profit of the manufacturer and the carbon emission. As Dominioni and Faure (2022) shows, an imbalanced policy can lead to either a loss of tax income or the limited effectiveness of emission policies. Although the government has spent a lot of time discussing the details of the policy and understands that even small parameter changes can affect policy effectiveness, evaluating the policy's effect is complex. Policies must precede market reactions, and past results are unreliable predictors due to changing demand and production situations. Once a policy is published, it is difficult to withdraw. For manufacturers, regardless of how the policy performs, they need to understand how different policies and changes in policy parameters will affect their production portfolio and revenue to adjust their business strategy. Different countries have their own goals in setting up policies, and it is essential for manufacturers to analyze how different policies in different countries differ to differentiate their production strategy and ensure better profitability.

Overall, this study aims to compare various emission policies and assess their impact on manufacturing portfolios for producers across different parameter scenarios in a quantitative manner. In this article, operations research methods are used to simulate the effect of different policies based on realistic test instances. By constructing mixed-integer linear optimization models for different policies and demand scenarios, policymakers as well as vehicle manufacturers can have a detailed quantitative view to assess policy effectiveness (Thies et al., 2022). The data comes from large global vehicle manufacturer, and the policy information is based on the current policy setup as of 2023. Moreover, since the mathematical model is flexible in changing parameters, the parameters and data can be adjusted to reflect the current situation and make more accurate predictions and analyses.

The remainder of this study is structured as follows: In Section 2, I discuss the literature related to this topic on finding out the emission policy impact. In Section 3, I present the model with detailed formulation and parameters, along with different models of the three emission policy systems in Europe, China, and the US. In Section 4, I show how the model is solved for different policy systems, with the main discussion focusing on the Super Credit System in Europe. In Section 5, I list out the test instances performed and the structure of the design of the experiment in evaluating different policy systems. In Sections 6, I present the final results, including the detailed portfolio, as well as sensitivity and factor analysis for different demand scenarios and parameter settings. Finally, in Section 7, I draw conclusions regarding the different policy systems and provide an outlook for further research.

2. Literature Review

Various approaches have been proposed to assess the effects of emission policies on car manufacturers, and they can be classified into five primary categories. Empirical studies and economic models rely on historical data and economic principles to conduct analyses on a broad scale. Market scenario models create hypothetical market conditions to evaluate policy impacts. A technology strategy model employs mathematical modeling to assist manufacturers in making decisions regarding the adoption of different vehicle models with various motive technologies. Simulation-based planning models use simulations to project long-term effects of the policy regulation for the manufacturer. Individual vehicle manufacturer mathematical models are tailored to specific manufacturers for in-depth analysis with output of detailed production plan and fleet emission trend.

In the category empirical studies and economic models, with the empirical studies gather data on emissions, produc-

tion, and market behavior to analyze the impact of emission policies and the economic model use economic principles to predict how policy changes may affect car manufacturers, such as changes in costs, prices, and market demand. For the empirical studies Bergek and Berggren (2014) reviewed the empirical studies on environmental policy and found that policy instruments play a crucial role in driving environmental innovation across sectors. Also, (Y. Wang et al., 2018) analyzed compliance strategies of four different automakers under dual-credit regulations, considering fuel economy and NEV (New Energy Vehicle) production which includes the ZEVs (Zero Emission Vehicles) and PHEVs (Plug-in Hybrid Electric Vehicle), comparing their approaches and suggesting cost-effective strategies with regulatory improvements.

Besides these empirical analyses, the economic model and pricing model have been built up to analyze how the policy would affect the vehicle manufacturer in a broad vision. (Moran et al., 2020) conducted micro-level studies with a multi regional input-output economic model to analyze the consumer-oriented policy and showed that these policies would reduce carbon emissions by about 25%. Additionally, the government pricing model of dual-credit policy published by (Yang et al., 2023), which compared with the market pricing model, shows that the dual-credit policy benefits energy saving and emission reduction in the transport sector. Moreover, in the study by (Ma et al., 2021), a supply chain model includes two stakeholders, the engine supplier and automakers, to analyze the carbon emission policy effect on the production of the ICEV and NEV vehicle. (Michalek et al., 2005) also considered the impact of the competition of other manufacturer in the paper and proposed mathematical models of engineering performance, consumer demand, and manufacturing costs, combined with game theory for the market segment. For these models, the trend could be observed from the economic perspective, but it could be too broad in scope that makes lack of some precision in explaining some details in the impact of the emission policy on the production plans of the individual manufacturer.

Besides the economic models, various market scenario models have been used to analysis the impact of the emission policy on the manufacturer's portfolio. These models creating various hypothetical market scenarios including the transportation sector and then evaluating how different emission policies would impact car manufacturers under these scenarios. For example, Thiel et al. (2016) used a TIMES-based energy system model to examines the impact of stricter EU CO2 car legislation on transport-related emissions, Electric vehicle uptake, oil consumption, and energy costs. This model is a modeling platform that consider factors like energy production, costs, and environmental impact and could helps make informed decisions about emission policies and resources. Hill et al. (2018) provided three models of PRIMES (global energy-economic model)-TREMOVE (transportation policy), GEM-E3T (model with macroeconomic, energy, and environmental policies) and the JRC DIONE (model for assessing energy and environmental policies) to analysis the overall market situation in considering energy, climate, transportation and the Europe emission policy. These models all reach similar conclusion that the EU policy are effective in reducing GHG(Greenhouse Gas) emissions. The ALTER-MOTIVE modelling method also been conducted by Ajanovic and Haas (2017) to integrate the energy system and transportation showing that GHG emissions could be reduced at least by 33% in a selected policy scenario. Other than these pre-formed model, Ou et al. (2018) develops the New Energy and Oil Consumption Credits Model to quantify the impacts of this policy in scenario from 2016 to 2020 to discuss the effect of the dual credit system in China on the electric vehicle sales. While these scenario models are useful for generating convincing results, they may oversimplify market conditions and the behaviors of individual manufacturers. These models can assess the effects of emission policies on a broad scale and from a market perspective, but they may not provide a comprehensive understanding of how individual car manufacturers would be impacted by or respond to these policies.

Speak to make analysis from individual level, there are some studies modeled the problem of individual car manufacturers to find profit maximizing technology strategies considering emission policies. S. Wang et al. (2018) build a mixed-integer mathematical model with decision variables representing various motor technologies in a technology combination (TC) problem. This model is designed to describe an automaker's decision-making process, and I utilize a genetic algorithm to assess the impact of China's dual credit policy from 2020 to 2025. Moreover, Romejko and Nakano (2017) increase the range of the motor technology path for the vehicle projects and explores a more diverse range of eight alternative fuel vehicles (AFVs), including EVs, FCVs, CNG (Compressed Natural Gas) vehicles, and more, to predict an optimal AFV portfolio for achieving economic and energy security goals. Moreover Zhu et al. (2022) proposed a decision-making algorithm for automakers' production strategies under the dual-credit policy in China. This algorithm reveals how automakers' choices transition between strategies based on thresholds and government targets. Furthermore, other than changing between the constraints and parameter, Kellner et al. (2021) also use multi-objective optimization method analysis in the technology selection to find the optimal power train technology portfolio. For these models, some individual level of the vehicle portfolio planning has been performed. However, these models only provide the outcome of which motor technology to initialize, without offering information on the actual quantity produced for different vehicle technology types. Consequently, they do not offer a comprehensive view of the resulting average fleet emission values or the detailed cost structure of vehicle manufacturers and are not capable for providing a very clear picture of the vehicle portfolio and the impact of emission policy on these portfolios.

Simulation method is also a popular modeling method to obtain vehicle portfolios. Kieckhäfer et al. (2012) at year 1970 used a hybrid market simulation approach for strategic planning of automotive vehicle portfolios to predict power-

train market shares across vehicle sizes based on portfolio offerings, consumer behavior, and market conditions. Also Kieckhäfer et al. (2009) creates a framework using system dynamics and agent-based simulation to analyze the product strategies which aid manufacturers in effective technology introductions across vehicle classes while considering regulations and markets. Moreover the NW (Newman and Watts) 'small world' network model also been used by Hu et al. (2020) to explore the dynamic effects of different policies on the diffusion of electric vehicles. For these simulation model, they could provide a close to reality output for the vehicle portfolio result thus indicate the impacts of the emission policy, but the model would take long times to be solved. It could be inefficient in time when implementing large number of testing scenarios into those simulation model for getting the solution insights.

Recently, Thies et al. (2022) proposed a novel model that concentrates on the individual vehicle manufacturing portfolio. This article analysis how the EU emission policy effects the vehicle portfolio using optimization model from the perspective of an individual vehicle manufacturer. This model has provided a detailed production portfolio, showing the optimal quantity of each vehicle type to be produced annually, along with a comprehensive resource plan and accurate average fleet emissions for each year's production and provide a basic framework of the optimization model used in this article. In this article, the base model of Thies et al. (2022) is extended by considering not only the EU policy but also the other two major policy system in China and the US. Furthermore, the EU policy is considered in more detail, as the super credit relaxation is considered. This novel model offers a more comprehensive and realistic analysis of policy effects from the perspective of individual car manufacturers and can be solved in few minutes. This model provides detailed information, including the vehicle initialization plan, production quantity for each vehicle type in each year, average fleet emissions in each year, market share of vehicle types, and the detailed cost structure. Moreover, it allows for easy parameter adjustments to customize the results as the policy changes.

3. Method

This section presents a mathematical model for project portfolio planning considering different emission policies. The overall road map for the model formulation is presented in Section 3.1 and Section 3.2 explains all the parameter information for the mathematical model. After that, a base model without emission policies is presented in Section 3.3, before it is extended by the emission policies for Europe, China and the U.S. in Section 3.4.

3.1. Model Road Map

Figure 1 describes the roadmap for the optimization model in this article. The planning horizon is 10 years, from 2025 to 2035, during which the manufacturer can change its portfolio production decisions. For the period outside the planning period, the settings remain fixed.

Within the planning horizon, the manufacturer can make several decisions, with the primary one being the determination of production quantities (q_{vt}) for each vehicle (v) in each period (t). To ensure these production quantities, the resource plan has also been established, which includes the capacity (k_{rt}) for each resource (r) in each period (t). Additionally, the resource adjustment plan contains (k_{rt}^{Rampup}) and ($k_{rt}^{Rampdown}$), representing the required increase and decrease amounts for resource (r) in period (t). The objective function aims to maximize the Net Present Value (NPV) for the entire production portfolio, which is discounted the manufacturer's profit obtained by subtracting all costs from the revenues. Revenues encompass sales revenue generated from vehicle sales throughout the planning period, as well as end-of-period capacity cash-out income. Costs include production expenses, which consist of both fixed and variable costs related to vehicle production, expenses for increasing or decreasing resources to maintain production capacity, development costs for initializing new vehicle projects, and penalty costs or gains dependent on the specific policy system in place.

Several constraints bind the decision-making process for vehicle production and can be categorized into five main categories. The vehicle project constraints help determine the initialization of different vehicle projects each year and limit the maximum number of projects that can be started. The production resource constraints are used to ensure sufficient resources are available for production. The vehicle demand constraint is employed to prevent the sale of more vehicles than the market demand, and the production volume constraint is used to ensure the minimum production volume each year. The policy-related constraints in different policy systems impose penalties or restrictions on the average fleet emissions of the production portfolio.

The inputs for modeling the production process constraints are depicted in Figure 1. The vehicle projects include several pre-defined projects categorized by powertrain technology type, size, production year, and power range class. Some vehicle projects are already determined before the planning horizon. Each vehicle project has its own production cost, sales revenue, and tailpipe line emission value. Development costs are assumed to be the same for all vehicle projects, and the maximum life cycle is equal for all projects. After 2025, within the planning period, the manufacturer can initialize new vehicle projects if suitable.Each vehicle project requires production resources. Before the planning period, the current on-hand resources are pre-defined. The manufacturer must ensure that resources can meet the production quantity of vehicle projects, which includes decisions on ramping up or down capacity. Resource costs include fixed and variable costs, with the fixed cost per production resource potentially decreasing due to economies of scale. Sales quantities must not exceed the market demand for a given year, segmented by vehicle type, size, and power range

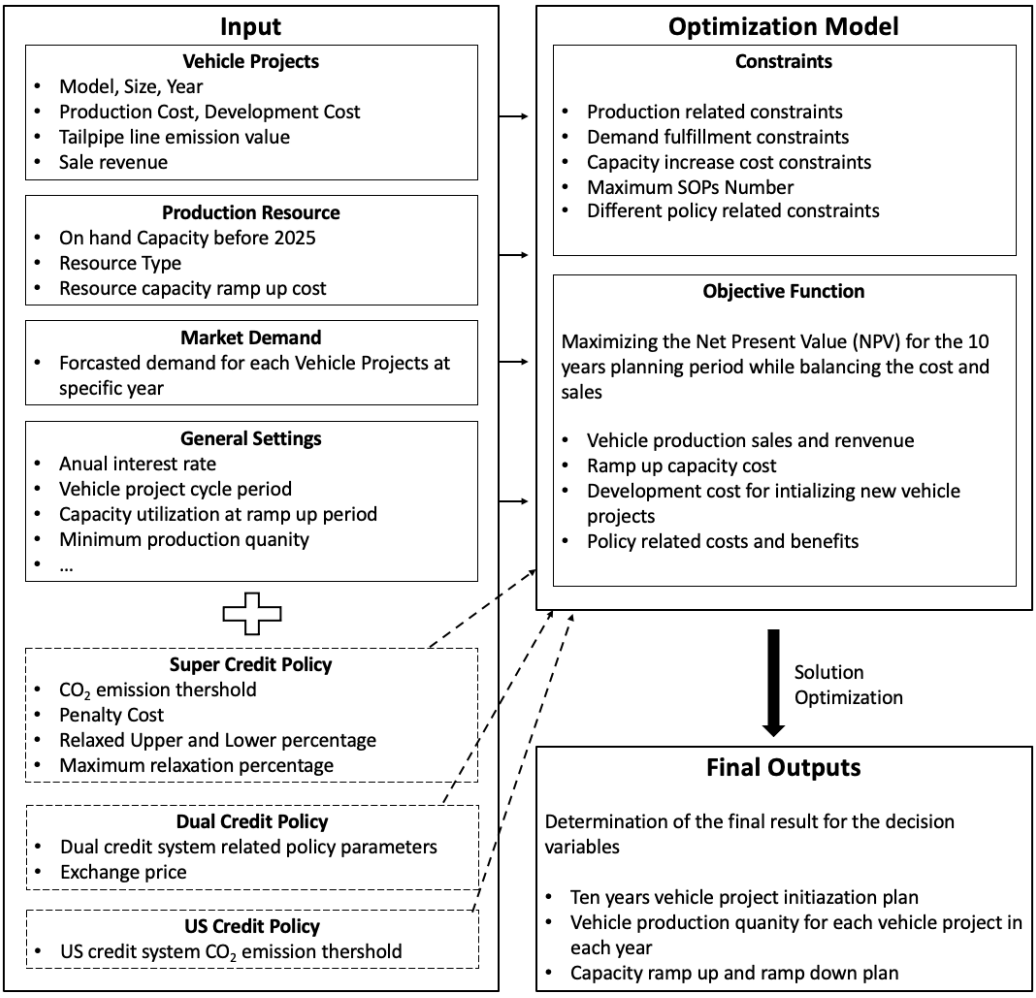


Figure 1: Optimization Model Road Map

(e.g., $ICEV_medium_low$). Multiple vehicle projects can be considered for several years within a market demand segment, but each vehicle project is associated with only one market segment. Quantities below market demand are assumed to be sold, and any unfilled demand is considered to be lost with no capacity payback. The general settings define the framework for general parameters, such as the assumed annual interest rate for NPV calculation, utilization loss due to capacity increment, and the minimum production quantity required to stay on the market, with full details described in section 3.2.3.

In addition to the basic setup, one of the policies from Super Credit Policy in Europe, Dual Credit Policy in China, US Emission Policy in the United States, or no Policy should be chosen to form the final policy-specific optimization model. Each policy entails specific parameters and formulas to be considered. The parameter values for different policies are based on current research in 2023, combined with personal assumptions, and all are related to the CO_2 average fleet emissions, as described in section 3.4.

Once these setups are incorporated into the model and the model is solved, the final objective value becomes available, along with detailed values for revenues and various cost sectors. Additionally, the fleet emissions for each vehicle project (v) in each period (t) can be determined by multiplying the carbon cycle emissions (E_v) by the quantity produced (q_{vt}) each year.

All details about the formulation of the optimization model are presented in section 3.3.

3.2. Parameter Information for the Vehicle Project Portfolio Planning

This section provides detailed parameter information, including sets, indices, decision variables, and descriptions for all parameters used in the model.

3.2.1. Sets & Indices

The model uses several sets and indices, which are described in Table 1. These sets and indices are essential for defining the parameters and decision variables in the model.

3.2.2. Decision Variables

There are 11 sets of decision variables described in Table 2. Three of these variables are binary variables, while

Table 1: Sets and Indices

Set & Indices		
Set	Index	Description
V	v	Vehicle projects
$X \in V$		Vehicle project start before the planning horizon
T	t	Periods
$P \in T$		Periods in planning horizon
M	m	Market segments
R	r	Production Resources
$E \in R$		Existing production resources at the beginning of the planning horizon

the rest are continuous variables. These decision variables play a critical role in the model, enabling the optimization of the vehicle portfolio and the assessment of various policy impacts.

3.2.3. Parameters

Table 3 concludes all the parameters used for the mathematical model, including the basic model as well as the additional emission policy model and separated by categories. The parameter contains type constant, variable, and vector.

3.3. Mathematical Formulation for the Base Model

This section describes the model for the base model without the emission policy but provides an explanation of the objective function and the constraints in the context of a pure vehicle manufacturer setting.

3.3.1. Objective Function

For the objective function, the goal is to maximize the net present value (NPV) of the vehicle project in ten years period. The interest rate is assume to be 5%, and the monetary value is calculated in each year and return the final NPV value base on year 2025. The objective function consists of three major parts which are the net profit of the production portfolio, the development cost and capacity increment cost, and finally the capacity cash back cost calculated and the end of the production planning phase at year 2035.

For the net profit of each year in the planning period, it adds up all sale revenue according to the production quantity and minus the production variable cost as well as the fixed cost, also the penalty cost due to the emission policy would be deducted according to different policy types.

For the development cost and capacity increment cost although the cost are spent on specific year in the production planning period, but it is assumed that the payment does not due immediately. The payment and deduction could be evenly distributed in 7 years period with each year about 14.3% of the total cost paid. $Vect_{vt}^{RD}$ is the parameter used for calculate the distribution of the cost and could also be counted after the production period. So the net NPV is summation for a all periods which is $t \in T$, from 2025 to 2050.

For the third part of the capacity back value, it is calculated at the end of the planning period at year 2035. This term used to count back the capacity value to prevent the model over estimated the cost for the last few years capacity increase. For the capacity back cost, it is assumed that the resource on hand would retain it's value in 10 years period and for each year, the value would depreciate by 10%. For example, for the resource in 2030, the resource would be capacity back with the rest value of 50%.

Objective Function:

Maximize NPV with:

$$\begin{aligned}\max \text{NPV} = & \sum_{t\in P}((TC_t^{SaleProd} - TC_t^{ProdFixed} - TC_t^{Penalty}) \cdot dr_t) \\ & - \sum_{t\in T}((TC_t^{RD} + TC_t^{Capacity}) \cdot dr_t) + TR^{CapaBack} \cdot dr_{T^{max}}\end{aligned} \tag{1}$$

with

$$TC_t^{SaleProd} = \sum_{v\in V}((s_v^{unit} - c_v^{Unitvar}) \cdot q_{vt}) \tag{2}$$

$$TC_t^{ProdFixed} = \sum_{r\in R}(c^{Resource} \cdot k_{rt}) \tag{3}$$

$$TC_t^{Penalty} = c_t^{Penalty} \tag{4}$$

$$TC_t^{RD} = \sum_{v\in V}(c_v^{RD} \cdot Vect_{vt}^{RD} \cdot y_v) \tag{5}$$

$$\begin{aligned}TC_t^{Capacity} = & \sum_{r\in R}\sum_{t\in P}(c^{RampUpFixed} \cdot Vect_{\tau t}^{PC} \cdot y_{rt}^{RampupBin} \\ & + c^{RampUpVar} \cdot Vect_{\tau t}^{PC} \cdot k_{rt}^{Rampup})\end{aligned} \tag{6}$$

$$TR^{CapaBack} = \sum_{r\in R} z_r^{RestValue} \tag{7}$$

3.3.2. Constraints

From the baseline model, there are four categories of constraint sets listed below. The constraints for different emission policies are in Section 3.4.

For the vehicle projects constraints set, it consists of constraints related to vehicle project initialization. Constraint (8) sets up the initial start for the vehicle project before 2025, the planning horizon. Constraint (9) uses a big number M to switch the binary variable for the start of the new vehicle project. Constraint (10) limits the allowed start for the number of vehicle projects in every year due to resource limits, and Constraint (11) forces the quantity produced to 0 if the vehicle exceeds the project life cycle of $t^{\max}$ years.

Table 2: Description of Decision Variables

Decision Variables			
Variable	Type	Range	Description
y_v	Binary	$\{0,1\}$	Vehicle project starting indicator with 1 meaning vehicle v is realized and 0 otherwise
q_{vt}	Continuous	$\mathbb{R}^+$	Number of vehicle project v produced at time period τ
k_{rt}	Continuous	$\mathbb{R}^+$	Capacity of resource r in period t
k_{rt}^{Rampup}	Continuous	$\mathbb{R}^+$	Increase of production resources r at the beginning of period t
$k_{rt}^{Rampdown}$	Continuous	$\mathbb{R}^+$	Decrease of production resources r at the beginning of period t
y_{rt}^{Rampup}	Binary	$\{0,1\}$	Resource increase indicator with 1 meaning there is an increase in resource r at the beginning of period t, with 0 otherwise
$z_r^{RestValue}$	Continuous	$\mathbb{R}^+$	Residence Value for resource r at the end of the planning horizon
$c_t^{Penalty}$	Continuous	$\mathbb{R}^+$	Penalty cost paid for the excess CO_2 Emission in period t
P_t^{relax}	Continuous	$\mathbb{R}^+$	Super credit policy relaxed percentage
y_1	Binary	$\{0,1\}$	Binary variable used to form maximum or minimum constraint
D_t^{Dual}	Continuous	$\mathbb{R}^+$	Cost paid(+) or Revenue earned (-) for dual credit policy

For the production resource constraints, Constraint (12) describes the resource usage constraint, and Constraint (13) indicates that in the ramp-up period, the resource would only be available at $\theta^{\max}$ percentage. Constraint (14) adjusts the on-hand resource at each year of the planning period after the previous ramp-up and ramp-down decisions that would be made. Constraint (15) switches on the fixed cost for ramping up the capacity. Constraint (16) calculates the rest value of the on-hand resources at the end of the planning year 2035, with V^{res} representing the remaining value, estimated as 5% of the total cost paid for increasing this amount of resource.

Constraint (17) is the vehicle demand constraint to ensure that the production volume, which is less than the demand, would be sold to earn profit. Constraint (18) guarantees the minimum production volume in each year in the planning period.

Vehicle Projects

$$y_v = y_v^{initial} \qquad \forall v \in X \quad (8)$$

$$q_{vt} \le M \cdot y_v \qquad \forall v \in X, \forall t \in P \quad (9)$$

$$\sum_{v \in V: SOP_v = t} y_v \le SOP^{\max} \qquad \forall t \in P \quad (10)$$

$$q_{vt} = 0 \qquad \forall v \in V, \forall t \in T : t \le SOP_v \vee t \ge SOP_v + t^{max} \quad (11)$$

Production resources

$$\sum_{v \in V: r_v = R} q_{vt} \le k_{rt} \qquad \forall r \in R, \forall t \in P \quad (12)$$

$$\sum_{v \in V: r_v = R} q_{vt} \le \theta^{max} \cdot k_{rt} \qquad \forall r \in R, \forall t \in P : t = t_r^{Rampup} \quad (13)$$

$$k_{rt} = k_r^{initial} + \sum_{\tau \in P: \tau \le t} (k_{r\tau}^{Rampup} - k_{r\tau}^{Rampdown}) \qquad \forall r \in R, \forall t \in P \quad (14)$$

$$k_{rt}^{Rampup} \le M \cdot y_{rt}^{Rampup} \qquad \forall r \in R, \forall t \in P \quad (15)$$

$$z_r^{RestValue} \le V^{res} \cdot \sum_{t \in U} ((y_{rt}^{Rampup} \cdot c^{RampUpFixed}) + (c^{RampUpVar} \cdot k_{rt}^{RampUp})) \qquad \forall r \in R \quad (16)$$

Vehicle Demand

$$\sum_{v \in V: m_v = m} q_{vt} \le d_{mt} \qquad \forall m \in M, \forall t \in P \quad (17)$$

Production volume

$$\sum_{v \in V} q_{vt} \ge q^{\min} \qquad \forall t \in P \quad (18)$$

Table 3: Detailed Parameter Information

Parameter			
Parameter Category	Parameter	Type	Description
Vehicle Projects	SOP_v	Vector	Start of production time for vehicle project v
	$c_v^{Unitvar}$	Constant	Unit variable cost for vehicle project v
	s_v^{Unit}	Constant	Unit sale revenue for vehicle project v
	E_v	Vector	CO_2 cycle emission of vehicle project v
	m_v	Vector	Market segment type of vehicle project v
	r_v	Vector	Resource type needed for vehicle project v
	t^{max}	Constant	Maximum duration of the selling time period
	$Vect_{vt}^{RD}$	Vector	Distribution of cash flow: Percentage of total development cost for vehicle project v in period t
	c_v^{RD}	Constant	Development cost for a new vehicle model
	$y_v^{initial}$	Vector	Preset vehicle realization indicator for vehicle project v
Production Resource	$c^{RampUpFixed}$	Constant	Fixed cost for ramping up production resources
	$c^{RampUpVar}$	Constant	Unit variable cost for the production resources
	t_r^{Rampup}	Vector	Ramp-up period for production resource r
	$Vect_{t_{cur}t_{rampup}}^{PC}$	Vector	Distribution of cash flow: Percentage of total production cost for ramp-up period t_{cur} distributed in current period t_{rampup}
	$c^{Resource}$	Constant	Constant cost for each unit of production resource on hand
	$k_r^{initial}$	Vector	Capacity of production resource r before the planning horizon
	V^{res}	Constant	Residence value for production resource at the end of planning horizon
Demand	d_{mt}	Vector	Number of vehicles demanded in market segment m at period t
	P_{min}	Constant	Minimum percentage of total demand needed to be fulfilled
Emission Regulation	E_t^{Law}	Constant	Total CO_2 fleet emission threshold in period t
	σ_{low}	Constant	The maximum fleet emission value for the category of low emission vehicles in the super credit policy system
	$\gamma_{mutiplier}$	Constant	PHEV multiplier factor in super credit policy system
	$E_t^{RelaxLaw}$	Variable	Relaxed threshold for total CO_2 fleet emission in period t
	$P_t^{EVthres}$	Variable	The regulated electric vehicle percentage to meet the relaxation criteria in period t
	$P^{SuperMax}$	Constant	Maximum allowance for relaxation in Super Credit System
	c^{CO_2}	Constant	Unit penalty payment cost per g/km for each vehicle sold
	S_t^{CAFC}	Variable	Total standard CAFC credit provided in period t
	A_t^{CAFC}	Variable	Actual CAFC credit consumed in period t
	S_{vt}^{FC}	Variable	Fuel consumption standard for vehicle project v in period t
	A_{vt}^{FC}	Variable	Actual fuel consumption for vehicle project v in period t
	W_v^{FC}	Variable	Weight factor for low emission vehicle project v in CAFC credit
	S_t^{NEV}	Variable	Total standard NEV credit for new energy vehicles required in period t
	A_t^{NEV}	Variable	Actual NEV credit gained for new energy vehicles in period t
	k_v	Variable	Weighted NEV credit ratio factor for vehicle project v
	R_t	Variable	Target ratio in period t discounted for NEV credit
	C^{Dual}	Constant	Monetary value for one dual credit in dual policy system

Table 3 — continued

Parameter			
Parameter Category	Parameter	Type	Description
Other Parameters	i	Constant	Assumed interest rate
	dr_t	Vector	Discount rate at period t
	q^{min}	Constant	Minimum production number required in each period
	$\theta^{\max}$	Constant	Maximum utilization of production resources in the ramp-up period
	SOP^{max}	Constant	Maximum number of vehicle project initializations in each period
Others	M	Constant	Large number used for modeling
	T^{max}	Constant	Last period in the planning horizon T

3.4. Mathematical formulation for Different Policy Systems

This section explains the three major emission policy models: the Super Credit System in Europe, the Dual Credit Policy in China, and the Emission Policy in the US.

3.4.1. Super Credit System in Europe

For the Super Credit System, the EU government has set a target tailpipe emission threshold to measure the average carbon emission each year. Furthermore, in Europe, there are not only simple penalty regulations for carbon emissions for vehicle manufacturers. Beginning in 2025, the EU Commission announced a Super Credit System on top of the CO_2 emission regulations. In simple terms, car producers can gain more CO_2 threshold reductions when they produce more low carbon emission vehicles. The definition of low carbon emission is when the CO_2 emissions of the car are lower than 50 g/km. With this Super Credit System, vehicle manufacturers would be more incentive to produce low or non-carbon emission vehicles since they would gain more CO_2 emission allowances or incur fewer penalty costs from a larger production of low carbon emission vehicles.

In the Super Credit System in Europe, the government calculates the percentage of low carbon emission cars in the whole car production portfolio. When the percentage of low carbon emission vehicles is calculated, the relaxed threshold can also be calculated. This threshold is relaxed as a percentage on top of the current year's CO_2 emission threshold. To calculate the percentage of the relaxed threshold, the percentage of the low carbon emission portfolio should first be compared to the regulated percentage set for low carbon emission vehicles for each year.

Mathematical Formulation for Super Credit System in Europe

The detailed model is listed below. Equation (19), with the help of Equations (20) and (21 3.21), represents the calculation of the relaxed threshold based on the production portfolio. Equations (20) and (21) set a lower limit with the help of the big number M, and Equation (22) sets the upper limit. Afterwards, the following equation incorporates the relaxed threshold percentage into the emission portfolio. This model will be further linearized for solving, and a detailed description of the linearized model can be found in Section 4.

$$\frac{\sum_{v\in\{V:E_v\le\sigma_{low}\}}(q_{vt}\cdot(1-\frac{\gamma_{mutiplier}}{\sigma_{low}}\cdot E_v))}{\sum_{v\in V}q_{vt}}\ge P_t^{relax}+P_t^{EVthres}-M\cdot(1-y_1)\qquad\forall t\in P\quad(19)$$

$$P_t^{relax}\ge -M\cdot y_1\qquad\forall t\in P\quad(20)$$

$$P_t^{relax}\le M\cdot y_1\qquad\forall t\in P\quad(21)$$

$$P_t^{relax}\le P^{SuperMax}\qquad\forall t\in P\quad(22)$$

$$E_t^{law}\cdot(1+P_t^{relax})\ge E_t^{Relaxlaw}\qquad\forall t\in P\quad(23)$$

$$\sum_{v\in V}(E_v\cdot q_{vt})\le E_t^{Relaxlaw}\cdot(1+\theta)\cdot\sum_{v\in V}q_{vt}\qquad\forall t\in P\quad(24)$$

$$z_t^{penalty}\ge(\sum_{v\in V}(E_v\cdot q_{vt})-E_t^{Relaxlaw}\cdot\sum_{v\in V}q_{vt})\cdot c^{CO_2}\qquad\forall t\in P\quad(25)$$

$$z_t^{penalty}\ge 0\qquad\forall t\in P\quad(26)$$

3.4.2. Dual Credit System in China

In China, the policy for encouraging the production of low-carbon emission vehicles differs from that in the EU. It is a dual-sided policy. Instead of setting a threshold with some relaxation criteria, the Chinese government has established a credit system for CO2 emissions. For a car production company, two different credit scores are calculated: Corporate Average Fuel Consumption (CAFC) and New Energy Vehicle (NEV) credits.

The CAFC credit score is more focused on the fuel consumption of the car portfolio, with a higher score for companies producing more traditional cars. The CAFC score has a target score calculated based on the current number of fuel consumption cars produced, as described in Equation (28). The coefficient is based on the size of the car and is calculated according to the vehicle's weight. The actual CAFC score is calculated based on the actual fuel consumption of the vehicle type, divided by a weighted factor based on the vehicle's powertrain technology type, as described in Equation (3.29). The final CAFC score is the difference between the standard CAFC score and the actual CAFC score.

For the NEV credit score, there is also a target score and actual score calculation, but the focus is on low-emission vehicles. The target score, which is also the standard NEV score, is calculated based on the quantity of ICEV cars with a coefficient of the target ratio for each year, as described in Equation (32). The actual NEV score is calculated based on the number of low carbon emission vehicles (PHEV, BEV, and FCEV) produced, with different coefficients for different types of vehicles, as described in Equation (31). Based on the policy regulated in 2020, the coefficients vary according to the model of the low carbon emission vehicle based on factors such as electric range multiplier (ER), battery energy density multiplier (BD), electric energy consumption multiplier (EC), and rated power multiplier (RP).

Once both the CAFC and NEV scores are obtained, the Chinese government requires car manufacturers to balance the two scores, which can be freely traded with banks or other companies. The modeling process for this exchange is complicated, so this model uses a monetary value to simulate the market exchange process. This means that car manufacturers could gain money as profit with more NEV credit scores earned and lose money with more CAFC scores.

Mathematical Formulation for Dual Policy in China

In this scenario, no additional constraints are required. However, for the objective function, the penalty cost would be replaced by the dual credit score, which consists of a CAFC credit score and a NEV credit score. The sum of these two scores would contribute to the final objective function, with a positive value indicating capital gain and a negative value representing a penalty cost. The monetary value of the score is based on the market price, but for simplicity in this model, the price of the score is assumed to be constant, based on past average values. This factor will be further discussed later.

CAFC SCORE

$$CAFC_t = S_t^{CAFC} - A_t^{CAFC} \qquad \forall t \in P \quad (27)$$

with

$$S_t^{CAFC} = \sum_{v \in V} (S_{vt}^{FC} * q_{vt}) \qquad \forall t \in P \quad (28)$$

$$A_t^{CAFC} = \sum_{v \in V} (\frac{A_{vt}^{FC}}{W_v^{FC}} * q_{vt}) \qquad \forall t \in P \quad (29)$$

NEV SCORE

$$NEV_t = A_t^{NEV} - S_t^{NEV} \qquad \forall t \in P \quad (30)$$

with

$$A_t^{NEV} = \sum_{v \in V:[BEV,PHEV,FCV]} (k_v * q_{vt}) \qquad \forall t \in P \quad (31)$$

$$S_t^{NEV} = \sum_{v \in V:[ICEV]} (R_t * q_{vt}) \qquad \forall t \in P \quad (32)$$

FINAL DUAL CREDIT SCORE

$$D_t^{Dual} = (CAFC_t - NEV_t) * C^{DualCredit} \qquad \forall t \in P \quad (33)$$

3.4.3. Credit System in US

In the United States, although there exists a CO_2 credit system, it is important to note that this system is only in place in the state of California. In the rest of the United States, there are federal vehicle emissions standards that car manufacturers must adhere to. These federal standards require car producers to ensure that their vehicle manufacturing portfolio strictly follows the restrictions for the average CO 2 emission of the fleet.

Mathematical Formulation for Emission Regulation in US

Under the U.S. emission law, an additional constraint (34) needs to be added to ensure that the carbon emissions for each year's production portfolio are strictly below the average emission standard set.

$$\sum_{v \in V} (E_v * q_{vt}) \leq E_t^{law} * \sum_{v \in V} q_{vt} \qquad \forall t \in P \quad (34)$$

4. Solution approach

This section present the solution approach for solving the optimization model. For the Chinese and US credit systems the optimization model is a mixed-integer linear program, which can be solved directly by a commercial solver. However, the EU credit system requires the non-linear constraints in equation (19), (24) and (25). Hence, to solve the model it must be linearized first. The proposed linearization approach is detailed in the following.

4.1. Linearization Approach

The original model for the super credit system in Europe introduces several non-linear constraints to the original model, making it unsolvable by standard solvers. The primary issue arises from the presence of the decision variable q_{vt} in both the numerator and denominator of constraint (19) in the original problem. Furthermore, constraints (24) and (25) involve the multiplication of the decision variable $E_t^{Relaxlaw}$ by q_{vt}. To solve this model, we need to transform the decision variable q_{vt}.

Using traditional methods, such as introducing new binary variables or using precewise linear function to replace the decision variable q_{vt} proves to be too complex for linearizing the model. This complexity arises because the decision variable q_{vt} appears in numerous instances throughout the model. Moreover, the decision variable q_{vt} exhibits a high degree of flexibility in terms of its range and value, as more than 120 decision variables must be determined over a 10 -year period. Achieving an optimal solution using conventional techniques would demand a significant amount of computational time. Since the primary goal of this paper is not to attain optimality, this paper employs an estimation approach instead. This approach allows me to find a heuristic solution that closely approximates the optimal solution, providing a practical resolution to the problem.

The approach functions in the following manner. Initially, I address the optimization problem without incorporating the relaxed policy, which solely consists of a threshold for fleet emissions and associated penalty costs. Subsequently, I reintroduce the relaxation aspect into the solution obtained in the previous step, thereby generating a heuristic solution for the European super-credit system. By utilizing this estimation approach, I can navigate the complexities introduced by the super credit system model and make informed decisions that align with the overarching objectives of the study. Detailed justification and implementation of this solution are presented in the following.

4.1.1. Mathematical Formula of the Linearized Model

In this linearized model, constraint (19) is redefined and substituted with constraint (35). The variable q_{vt} in the denominator is replaced with $P_{\min} * \sum_{m\in M} d_{mt}$, where $P_{\min}$ is an adjustable parameter used to estimate the total production volume. Additionally, constraint (42) is introduced to ensure that the total production volume exceeds the estimated production quantity in the planning period after 2025, as the capacity is fixed starting from the year 2025 and cannot be altered. The replacement of the decision variable q_{vt} also applies to constraints (24) and (25), which are replaced by constraints (40) and (41). Through this approach, q_{vt} is transformed into a parameter that can be adjusted using P_{min}, making the problem solvable for the purpose of testing the results. The detailed formulation of the model is described below.

$$\frac{\sum_{v\in\{V:E_v\le\sigma_{low}\}}(q_{vt}\times(1-\frac{\gamma_{mutiplier}}{\sigma_{low}}\times E_v))}{P_{min}\cdot\sum_{m\in M} d_{mt}} \ge P_t^{relax} + P_t^{EVthres} - M\cdot(1-y_1) \qquad \forall t\in P \quad (35)$$

$$P_t^{relax} \ge -M\cdot y_1 \qquad \forall t\in P \quad (36)$$

$$P_t^{relax} \le M\cdot y_1 \qquad \forall t\in P \quad (37)$$

$$P_t^{relax} \le P^{SuperMax} \qquad \forall t\in P \quad (38)$$

$$E_t^{law}\times(1+P_t^{relax}) \ge E_t^{Relaxlaw} \qquad \forall t\in P \quad (39)$$

$$\sum_{v\in V}(E_v\cdot q_{vt}) \le E_t^{Relaxlaw}\cdot(1+\theta)\cdot P_{min}\cdot\sum_{m\in M} d_{mt} \qquad \forall t\in P \quad (40)$$

$$z_t^{penalty} \ge (\sum_{v\in V}(E_v\cdot q_{vt}) - E_t^{Relaxlaw}\cdot P_{min}\cdot\sum_{m\in M} d_{mt})\cdot c^{CO_2} \qquad \forall t\in P \quad (41)$$

$$\sum_{v\in V} q_{vt} \ge P_{min}\cdot\sum_{m\in M} d_{mt} \qquad \forall t\in P_{2025} \quad (42)$$

$$z_t^{penalty} \ge 0 \qquad \forall t\in P \quad (43)$$

4.1.2. Evaluating the Quality of the Approximation

By performing this replacement, the objective function becomes prone to overestimation, as the new variable $P_{\min} * \sum_{m\in M} d_{mt}$ will always be smaller than the actual production quantity q_{vt}. The parameter $P_{\min}$ is constrained not to fall below 70% because it is essential to maintain a reasonable production volume, and excessively low production quantities are impractical. The gap between the overestimated result and the real result can be analyzed and compared in order to search for a heuristic solution that closely approximates the optimal solution. If the estimation leads to a local minimum with a lower gap and a high objective value, it suggests that the optimal solution might fall within this production quantity range.

For the gap analysis, P_{min} is tested within a range from 70% to 100%, with intervals of 5%. The model is initially solved by setting the $P_{\min}$ value to obtain the real production quantity. Subsequently, the actual production quantity is reintroduced into the original problem to calculate the actual NPV value. Both results are collected and recorded for further analysis.

The observed gap is approximately 2%, and it shows relatively little fluctuation. Additionally, a peak objective result is observed at the point where the $P_{\min}$ value is set to 90% with the lowest gap. Detailed graphical representation can be found in Appendix (see Figure A.1). Based on these results, it is reasonable to assume that the optimal solution may fall within this range, but a more precise estimation is required.

For this refined estimation, the testing range is narrowed to 90% to 100% but with a higher interval frequency of 1%.

A detailed graph has been generated, and according to this graph (see Figure A.2), 91% is the estimated point with the highest objective value and the lowest gap. However, using this estimation method, the quantity produced in each year of the planning period is nearly identical, making it challenging to perform tests to find the exact optimal solution. To interpret the 91% estimation, a relaxed model including the Europe emission policy but excluding the Super credit relaxation needs to be established. For this model, the constraint of the super credit policy is excluded, and the additional emission related constraints used is described below with only the fleet emission threshold along with the penalty cost.

$$\sum_{v\in V}(E_v \cdot q_{vt}) \leq E_t^{law} \cdot (1+\theta) \cdot \sum_{v\in V} q_{vt} \qquad \forall t \in P \quad (44)$$

$$z_t^{penalty} \geq (\sum_{v\in V}(E_v \cdot q_{vt}) - E_t^{law} \cdot \sum_{v\in V} q_{vt}) \cdot c^{CO_2} \qquad \forall t \in P \quad (45)$$

$$z_t^{penalty} \geq 0 \qquad \forall t \in P \quad (46)$$

This linear model can be solved, and the results indicate that the optimal average output production quantity over a ten-year production horizon is also around 91%, which closely aligns with the estimated $P_{\min}$ value. Furthermore, when the credit relaxation is applied back to the resulting production quantities, the objective value increases by approximately 2% with a higher net NPV value. These results suggest that this revised model, excluding the super credit relaxation, could serve as a starting point for calculating a satisfactory heuristic solution. To assess the reliability of these results, the same analysis is repeated for a different demand scenario, assuming that the market is more innovative and people are more receptive to low carbon emission vehicles. The detailed result graphs are provided in Appendix A. 3 and A. 4.

The results show that the production quantity at approximately 86% corresponds to the highest objective value and the lowest result gap, with a difference of about 1.67%. For the revised model without the super credit relaxation, the optimal production quantity is around 84%, which is also close to the estimated result. It also yields the highest objective value, at a level of approximately 2% higher than the estimated result.

4.1.3. Approximation Approach used for Experiments

In conclusion, the revised model, which excludes the super credit relaxation, has undergone testing for reliability and sensitivity by varying the production quantity within a range of -5% to 5% in both conservative and innovative demand scenarios. Subsequently, the super credit relaxation policy was reintroduced to obtain the final results. The detailed graphs in the Appendix (see Graph A. 5 and A.6) demonstrate that this revised model consistently yields the largest objective value locally.

As a result, it can be concluded that solving the super credit system model can be performed in two phases. First, utilize the model without the relaxation policy as the initial estimation. Then, reintroduce the relaxation policy to obtain the final heuristic result. While this result may not be optimal, it is sufficiently close to the optimal solution for further analysis.

5. Design of Experiment

This section outlines the experimental design for the computational study of the different emission policy model. In Section 5.1, the foundational data for modeling the system under distinct policy frameworks is introduced. Sections 5.2 and 5.3 detail the procedures used to analyze the the impact of the different policies and their parameters.

5.1. Test Instances

First, in Section 5.1.1, I present a detailed description of the data and specifications for the base model. This base model excludes the effects of the emission policy system and provides specific numerical values for various parameters based on personal assumptions and research. Second, in Section 5.1.2, I provide specific numerical values for different parameters within various emission policies in Europe, China, and the U.S.

5.1.1. Test Instances for the Base Model

The data originates from a passenger car manufacturing company in Europe. It is derived from publicly available data sets with same source used in the Thiel et al. (2016) paper, supplemented with certain assumptions and settings based on personal research.

Vehicle Projects

A total of 264 vehicle projects have been defined spanning from the year 2025 to 2035. These vehicle types are categorized based on powertrain technology (ICEV, PHEV, FCEV, BEV). Furthermore, within each powertrain technology category, they are further classified by power range, including 'small,' 'medium,' and 'large,' as well as discrete power levels, denoted as 'high' or 'low.' It's important to note that FCEV and BEV vehicles are considered zero-emission vehicles, while PHEVs are classified as low-emission vehicles. Additionally, for ICEV vehicle types, it is assumed that advancements in technology will lead to a reduction in their pipeline emissions, as is the case for PHEVs. Conversely, it is assumed that variable production costs for ICEVs and PHEVs will increase in the future.

The sale revenue for each vehicle project varies depending on the year, size, and powertrain technology, falling within a range of € 8,400 to € 28,000. Furthermore, each

vehicle project is assumed to have a life cycle of seven years. The development cost for each vehicle project is assumed to be uniform and approximately € 420 million. More detailed information can be found in the Data Information Excel File.

Production Resources

For different discrete power levels ('high' or 'low'), the same set of resources can be utilized. In terms of production resources, the capacity required to meet the production plan can be increased annually. However, increasing capacity entails a fixed cost of €20 million, and each incremental unit of capacity incurs a variable cost of € 2.75 as part of the ramp-up process. It's important to note that during the rampup period, only 75 % of the total capacity for that year can be effectively utilized. In addition to production variable costs, there is also an additional fixed cost of € 50 associated with each vehicle produced.

Vehicle Demand

The total assumed demand for vehicles is 2.25 million units per year, spanning from 2025 to 2035. It's important to note that the total demand remains constant over this ten-year planning period. However, the market share for the four types of vehicles will vary. Each year, the percentage split for each powertrain technology is assumed to remain unchanged. In this model, I consider two distinct demand scenarios categorized into two main groups: 'innovative' and 'conservative.' Figure 2 illustrates these two demand levels.

In both scenarios, ICEVs dominate the market share, starting at about 75 %, and gradually decrease their market share as PHEVs, BEVs, and FCEVs gain ground. In the 'innovative' scenario, I have a more optimistic outlook on electric cars (BEVs and FCEVs) and assume that the market will be more accepting of them. As a result, the market share for these powertrains increases linearly each year at a higher rate compared to the 'conservative' scenario. Conversely, in the 'conservative' scenario, I assume that people will be more cautious in adopting purely electric cars, so the market share of PHEVs may increase more rapidly.

Overall Manufacturing Settings

In this model, the objective value NPV was evaluated at a discount factor of 5% each year, and the final objective function is based on the year 2025. Additionally, for the allocation of the increased capital expenditure, it is assumed that the spending is evenly distributed over 7 years. Furthermore, the vehicle projects are determined in the years leading up to 2025, specifically from 2020 to 2024, and the capacity can only be changed in 2025. All constant parameters are summarized in Table 4, with further data available in the Data Information Excel File.

Table 4: Constant Parameter Values

Constant Parameters Value		
Parameter	Unit	Value
t^{max}	years	7
c_v^{RD}	€	420 Million
$c^{RampUpFixed}$	€	20 Million
$c^{RampUpVar}$	€ / (car × year)	2.75
$c^{Resource}$	€ / (car × year)	65
i	%	5
θ^{max}	%	95
SOP^{max}	vehicle projects	15
q^{min}	cars	1 Million

5.1.2. Test Instances for the Different Policy System

The data is sourced from regulatory documents. This information is presented in different sections dedicated to the different policy systems.

Super Credit System

For the test instance used in modeling the super credit system, the information is sourced from the official website of the European Union (European Commission, 2023). The emission threshold from 2025 to 2029 is set at approximately 80 g/km, and by 2030, it is reduced to 60 g/km. Further reductions occur by 2035 when the threshold is set at 45 g/km. In the event of exceeding the carbon emission limits, the government would impose a penalty cost of € 95 per unit of excess emissions (g/km) for each vehicle sold.

For the relaxed percentage of the low carbon emission vehicle, the current regulation stipulates that from 2025 to 2029, car producers must exceed a minimum of 15 % in low carbon emission vehicles to qualify for the relaxed threshold. From 2030 to 2035 , this percentage requirement increases to 35 %. Additionally, the maximum allowed increase for the threshold relaxation is 5 % per year. For example, in 2030, if a car manufacturer achieves a 38 % production of low carbon emission vehicles, the relaxed threshold percentage would be 3 %. If the low carbon emission vehicle production further increases to 45 %, the manufacturer would receive the maximum 5 % relaxed percentage. Table 5 provides a summary of the parameters for the super credit system, and the mathematical formulation of the super credit system is detailed in the section below. Also when calculating the percentage of low carbon emission vehicles, the coefficient of low emission vehicles in the numerator is linearly dependent on the actual carbon emissions. With the inclusion of the PHEV multiplier in the current policy setting, vehicles with 50 g/km carbon emissions would start with a coefficient of 0.3 and increase linearly as carbon emissions decrease. For example, a zero-emission vehicle would have a coefficient of 1 , representing

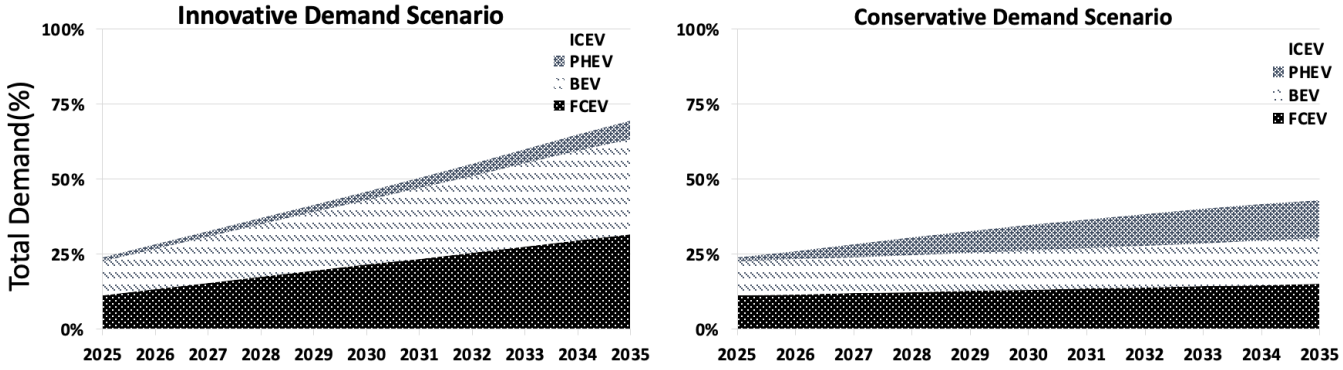


Figure 2: Vehicle Demand

Table 5: Super Credit System Parameters in Europe

Super Credit System Parameter Values			
Factors	Symbol	Unit	Value
Emission threshold	E_t^{law}	g/km	80-60-45
Fleet emissions limit for low emission vehicle	σ_{low}	g/km	50
Penalty cost	c^{CO_2}	€ / (g/km)	95
Relaxed threshold before 2030	$P_t^{EV\,thres}$	Percentage	15%
Relaxed threshold after 2030	$P_t^{EV\,thres}$	Percentage	35%
Maximum relaxed threshold limit	$P^{SuperMax}$	Percentage	5 %
PHEV multiplier	$\gamma_{mutiplier}$	-	0.7

one car produced, while a car with 25 g/km emissions would have a coefficient of 0.65.

Dual Credit System

For the test instances of the dual policy system in this model, the policy information is referenced from a report by the ICCT (International Council for Clean Transportation) (International Council on Clean Transportation (ICCT), 2016).

Regarding the target values for the CAFC score represent in Equation (28), fuel consumption vehicles are categorized into three groups (low, medium, large). The standard values S_{vt}^{FC} are determined by applying the weight of each vehicle size to the Chinese government's fuel consumption standard formula (TransportPolicy.net, 2018). Furthermore, after 2030, the Chinese government plans to further reduce the standard from an average of 4.0 $L/100\ km$ to 3.2 $L/100\ km$, resulting in lower fuel consumption standards for vehicles of different sizes (International Energy Agency (IEA), 2021). The resulted final consumption standards are detailed in Table 6.

For the actual CAFC score calculation represent in Equation (29), the value is converted from the carbon emission value. The actual coefficient A_{vt}^{FC} for fuel consumption is expressed in units of $L/100\ km$. The conversion from g/km to $L/100\ km$ is given by 100 $g/km = 4.25\ L/100\ km$. Additionally, since the data is sourced from an EU country and the measurement system differs, the Chinese government would consider about a 15% adjustment (Tietge et al., 2017). Therefore, the final actual fuel consumption should be multiplied by a factor of 1.15. Equation (5.1) provides a simple example of the conversion calculation. Furthermore, for the weighted factor W_v^{FC} in calculating the actual CAFC score, it is set to 1 for all types of vehicles in accordance with current regulations.

Table 6: Fuel Consumption Standard for Different Vehicle Sizes (2025-2035)

Fuel Consumption Standard (2025-2035) [$L/100\ km$]			
Size	Weight	Before 2030	After 2030
Small	680 kg	3.6	2.88
Medium	1500 kg	4.05	3.22
Large	1995 kg	4.398	3.518

$$100(g/km) = 4.25 \times 1.15 = 4.8875(L/100km) \quad (47)$$

For the target NEV score calculation is represent in Equation (32). As for the target ratio R_t, it can vary from year to year. In 2025, the target ratio is approximately 20%, increasing to 40% by 2030, and eventually reaching 50% by the end of 2035. The target ratio can be adjusted each year, although there are no official rules announced. In this model, the target ratio is assumed to increase linearly and evenly each year, with specific percentages detailed in Table 7.

Table 7: NEV Target Ratios for the Years 2025-2035

NEV Target Ratio (2025-2035) [%]	
Year	Target Ratio
2025	20%
2026	24%
2027	28%
2028	32%
2029	36%
2030	40%
2031	42%
2032	44%
2033	46%
2034	48%
2035	50%

The formula for the calculation of the actual NEV score is in Equation (31). The weighted factor k_v for the calculation of the actual NEV score could be hard to predict since the value is different for specific vehicle model, to simply the model, this model use the average score value and shown in table 8.

Table 8: NEV Credits per Vehicle Type

Weighted Factor per Type	
Vehicle Type	Estimated NEV Credit
PHEV	1.6
BEV	4
FCEV	4.8

For the monetary value $C^{DualCredit}$, in Chen and He (2022) article, the average exchange cost for one credit is between $2600-2900$ RMB, so in this model it is assumed to be 330 Euro after the currency exchange.

U.S. Emission Policy

The detailed emission threshold values E_t^{law} in Equation (34) used in the US emission model are sourced from the standard released by the US Environmental Protection Agency (Register, 2023) and are summarized in Table 9. To maintain consistency and precision, the units in the policy, originally given in units of $g/mile$, have been converted to units of g/km, with values rounded to one decimal digit.

5.2. Policy Comparison Experiment

This paper aims to employ mathematical models to explore the differences among three distinct emission policies in Europe, China, and the US, while comparing them to the baseline model with no emission policy in place. The results will be evaluated under two different demand scenarios: conservative and innovative demand. Each policy's parameters have been calculated and estimated based on publicly available information.

For each policy scenario and demand type, a comprehensive analysis will be conducted, including an examination of the vehicle initialization schedule and production quantities. Additionally, beyond the decision variables, the objective function will be thoroughly examined. This examination will encompass cost structures and average fleet emissions to gain insights into the impact of various emission policies. Table 10 summarizes the experiment's outline.

5.3. Sensitivity and Factorial analysis for Different Emission policies

In addition to comparing different emission policies, this paper will also conduct sensitivity and factor analyses for each emission policy to assess how policy parameters affect policy effectiveness and identify the parameters with the most significant impact on emission policies. For sensitivity analysis, each factor will be categorized into three levels: low, basic, and high, and tests will be conducted at these levels for analysis. In the factorial analysis, a 2-level factorial analysis will be performed, reducing each factor to two levels: high and low. A $1/2$ fraction of the full factorial design method will be used to enhance the efficiency of the factorial analysis.

For the analyzed outputs, due to the extensive testing required, only two types of outputs will be compared: the net total NPV (Net Present Value) and the percentage of low carbon emission vehicles. NPV is the objective function favored by car manufacturers, as they seek to maximize NPV. However, this objective may conflict with the the reduction of carbon emissions. Therefore, the percentage of low carbon emission vehicles will also be analyzed to assess the policy's impact on social benefits from the government's perspective and evaluate policy effectiveness. Table 11 summarizes the Design of Experiment (DOE) for the emission policy.

5.3.1. Super Credit System

For the Super Credit system policy in Europe, four factors will be considered for the analysis. The first factor is the demand type, which includes two scenarios: conservative and innovative. The next factor is the threshold percentage for relaxation, which is defined as the percentage of low emission vehicles with less than 50 g/km emissions. The threshold is set in two stages, one before 2030 and one after 2030.

Table 9: Federal Vehicle Emissions Standards

Year								
	2025	2026	2027	2028	2029	2030	2031	2032 and later
Emission standard (g/km)	149	152	134	116	99	91	82	73
Emission standard (g/km)	92.9	94.8	83.5	72.3	61.7	56.7	51.1	45.5

Table 10: Comparison of Different Policy Experiments

Policy Comparison Experiment		
Policy Type	Demand	Output
Super Credit System	Con/Inno	• Vehicle initialization schedule • Vehicle production quantity • Cost structure • Average fleet emission result
Dual Credit System	Con/Inno	
US Credit System	Con/Inno	
Baseline Model (No emission policy)	Con/Inno	

For the sensitivity analysis, the range of the threshold before 2030 is from set at level 5%, 15% and 25%, and after 2030, it is at the level of 25%, 35% and 45%. The third factor is the maximum allowed threshold percentage, which is at the level of 1%, 5% and 9%. Finally, the PHEV multiplier, which provides a base percentage for PHEV-type vehicles, is included in the analysis. It can be set to either "on" or "off" to test its effect on NPV value and the actual quantity of low carbon emission vehicles. These factors will be analyzed to understand their impact on the NPV value and the quantity of low carbon emission vehicles.

5.3.2. Dual Credit System

In the China dual credit system, four factors will be considered for analysis. The first factor is the demand type, which is similar to the super credit system. The remaining three factors are tested by percentage changes, and they include one important indicator for calculating the CAFC score, one indicator for the NEV score, and one factor considering the exchange cost for credit score realization. The level of change for these factors is the same which are -50%, 0% and 50%. These factors will be analyzed to understand their impact on the NPV value and the quantity of low carbon emission vehicles.

5.3.3. US Credit System

For the US credit system, apart from the demand type factor, one more factor which is the percentage change of regulated threshold are considered at level of -10%, 0% and 10%. Also strict or non-strict compliance would be tested as another factor in this scenario. For strict compliance meaning that the threshold could not be exceed and non-strict compliance it is assumed the penalty cost would be similar in EU with 95 €/((g/km) × year).

6. Result

To assess the impact of various policy systems, Section 6.1 includes an analysis of the optimal production plan under four different policy scenarios. Additionally, the consideration of a multitude of parameters comes into play, with each having the potential to influence the final outcomes of these regulations. For this purpose, Section 6.2 conducts sensitivity analysis at three different levels (low, basic, and high) for selected parameters under different policies, while Section 6.3 performs a factor analysis of these parameters.

6.1. Optimal Production Plan

The detailed plan includes the initialization of vehicle projects for each year from 2025 to 2035 , as well as the production quantity for four different types of powertrain technologies.

In the vehicle project initialization section, the blocks are separated by powertrain technology, vehicle project size, and year. Gray blocks represent the start of a powertrain technology in that year, while white blocks indicate that the vehicle project will not be initiated. For the production quantity, each year from 2025 to 2035 is categorized by powertrain technology and filled with different patterns, as described in the stacked bar chart.

Additionally, the average fleet emissions for each year were plotted on a line chart, with the EU-regulated fleet emission threshold as a reference for comparison. The objective value, composed of six main components [Capacity income at the end of the period, Capacity increase cost, Development cost, Profit, Fixed cost, Penalty cost/Dual Credit value], was depicted in a bar chart. The final Net Present Value (NPV) was shown on a line chart for further analysis. These visualizations provide insights into the different policy scenarios and their effects on production planning and emissions.

Table 11: Summary of Design of Experiments (DOE) for Emission Policy

Summary for DOE of Emission Policy				
Policy Type	Factors	Level	Type	Outputs
Super Credit System	Demand Type	ConInno	Option	NPV, EC percentage
	Relaxed Threshold	Before 2030 [5% or 25%] After 2030 [25% or 45%]	Percentage	
	Maximum Threshold	1% or 9%	Percentage	
	PHEV Multiplier	Y/N	Option	
Dual Credit System	Demand Type	Con/Inno	Option	NPV, EC percentage
	% change of Standard Fuel Consumption	-50% or 50%	Percentage	
	% change of exchange price	-50% or 50%	Percentage	
	% change of NEV weight factor	-50% or 50%	Percentage	
US Credit System	Demand Type	Con/Inno	Option	NPV, EC percentage
	% change of CO_2 threshold	-10% or 10%	Percentage	
	Strict compliance	Y/N	Option	

Table 12: Parameter Changes for Dual Credit System

Dual Credit System Parameter Changes			
	Standard Fuel Consumption (CAFC)	Exchange Price	NEV Weighted Factor (NEV)
Example Base (small_PHEV_low_2025)	3.6	€330	1.6
-50% Percentage Change	1.8	€165	0.8
+50% Percentage Change	5.4	€495	2.4

Table 13: Parameter Changes for US Credit System

US Credit System Parameter Changes								
Year	2025	2026	2027	2028	2029	2030	2031	2032 and later
Emission standard (g/km)	92.9	94.8	83.5	72.3	61.7	56.7	51.1	45.5
-10% Percentage Change	83.6	85.3	75.2	65.1	55.5	51.3	36.0	41.0
+10% Percentage Change	102.2	104.3	91.9	79.5	67.9	62.4	56.2	50.1

6.1.1. Conservative Demand

The following results were calculated under the assumption of a conservative market demand for low-emission vehicles, characterized by a lower growth rate in low-emission vehicle adoption. This scenario represents a more cautious market approach towards low-emission vehicles.

Detailed Portfolio Analysis

Figure 3 presents information for the four different emis-

sion policies. In the scenario of no emission system, which can be considered a reference point for interpreting the other three policies, there are no penalties or restrictions on excess fleet emissions. In this scenario, only vehicle projects that contribute positively to the net present value of the portfolio are initiated. From this scenario, it is evident that all types of ICEV vehicles in each year have a positive effect on the profit margin. However, for the small-sized PHEV, FCEV, and BEV vehicles, the profit margin is consistently negative, so there is no incentive for manufacturers to produce these vehicle projects. For medium and large-sized vehicles in these powertrain technologies, as production costs are assumed to decrease due to technological development, medium and large-sized PHEV projects are initiated after 2029, medium-sized FCEV projects after 2031, large-sized FCEV projects after 2029, and all medium and large-sized BEV projects are initiated due to their positive profit margins.

After analyzing the base scenario with no impact from emission policies, three different major policies can be analyzed. Figure 3 shows that several vehicle types with negative profit margins are initiated in order to balance penalty costs or meet the constraints on excess fleet emissions. The Dual Credit Policy results in the lowest number, with about 70 initiated low-emission vehicle projects during the ten-year planning period. It is followed by 76 vehicle projects in the Super Credit System, while the US Credit System leads with the highest number, with about 87 initiated projects. ICEV vehicles dominate the portfolio, but in the US Credit System, due to strict emission standards after 2029, small-sized ICEV vehicles are terminated because of lower profit margins compared to medium and large-sized ICEV vehicles. For PHEV types, in the US Credit System, several small-sized PHEV projects are initiated to meet emission constraints. However, in the Super Credit System and Dual Credit System, no small-sized PHEV projects are initiated. Furthermore, at the end of 2035, in the Super Credit System and US Emission System, there is a sudden drop in the number of large-sized PHEV projects, likely due to the sufficient production of zero-emission vehicles. For these projects, the profit margin decreases, potentially falling below that of zero-emission vehicle types due to increased production costs. For FCEV and BEV types of vehicles, medium and large-sized versions are initiated to balance the high CO2 emissions from ICEV vehicles, and small-sized versions are initiated as a last resort. In comparison, BEV projects are more favorable for manufacturers due to their higher profit margins.

Regarding production quantities, it is evident that manufacturers tend to align with market demand and produce as many ICEV vehicles as possible. Comparing the results to the base model with no emission policy, it can be observed that all three emission policies would reduce the production quantity of ICEV vehicles, with the US Emission Policy leading to the largest reduction. In the Super Credit System, there is a sudden drop in the production of ICEV vehicles at year 2035. This drop is a result of the fleet emission threshold decreasing from 60 g/km to 45 g/km, causing manufacturers to reduce production to meet the stricter standards. For the Dual Credit Policy, the initial production quantity in 2025 is lower due to lower demand for low-emission vehicles and the lower fuel efficiency of ICEV vehicles. However, in the following years, production quantities increase since the ICEV vehicle become more fuel efficient thus turns out to become more favorable in the dual credit score system.

Fleet Emission and NPV Analysis

The fleet emissions and objective function composition for the four different policy scenarios are depicted in Figure 4. The red line represents the EU-regulated threshold for average CO_2 fleet emissions and serves as a reference to assess the impact of different emission policies. In the base model, shown by the black line, there is some decrease in fleet emissions due to technology development, but the average CO_2 fleet emissions consistently remain above the EU threshold meaning some regulation need to be performed as an external force to control the fleeting emission. The three different emission policies all have some effect on reducing fleet emissions in each year. In the Super Credit System (blue line), fleet emissions follow the EU-regulated threshold with some relaxation in certain periods. In the Dual Credit System (yellow line), the trend is similar to the line with no emission policy but with smaller absolute fleet emissions. However, fleet emissions are still above the EU-regulated threshold. Lastly, in the US Emission System (green line), fleet emissions strictly follow the US emission threshold, which becomes more stringent after 2028.

In terms of net present value, the base scenario yields the highest number. In the Dual Credit Policy, the NPV is also high because manufacturers can earn money for producing low-emission vehicles. Under conservative demand, manufacturers would decide to produce more low-emission vehicles to earn these new vehicle credits that could be trade to earn some money. In the Super Credit System, a minor amount of penalty cost is incurred, resulting in a 12% decrease in the net present value compared to the base model. The US Emission System yields the lowest net present value, as targets must be strictly met, leading to the production of several low-profit-margin vehicle types and a 25% decrease in the objective value.

6.1.2. Innovative Demand

In the Innovative demand scenario, the market is more receptive to low-emission vehicles, with a higher growth rate in the low-emission vehicle market.

Detailed Portfolio Analysis

In this demand scenario, the market introduction of new vehicle project is similar but with some minor differences, as described in Figure 5. In the baseline situation, medium-sized FCEV projects in 2030 would also be initiated due to higher demand, and the large-sized BEV vehicle project in 2025 would not be started to prioritize the production of ICEV types. Compared to the conservative demand scenario,

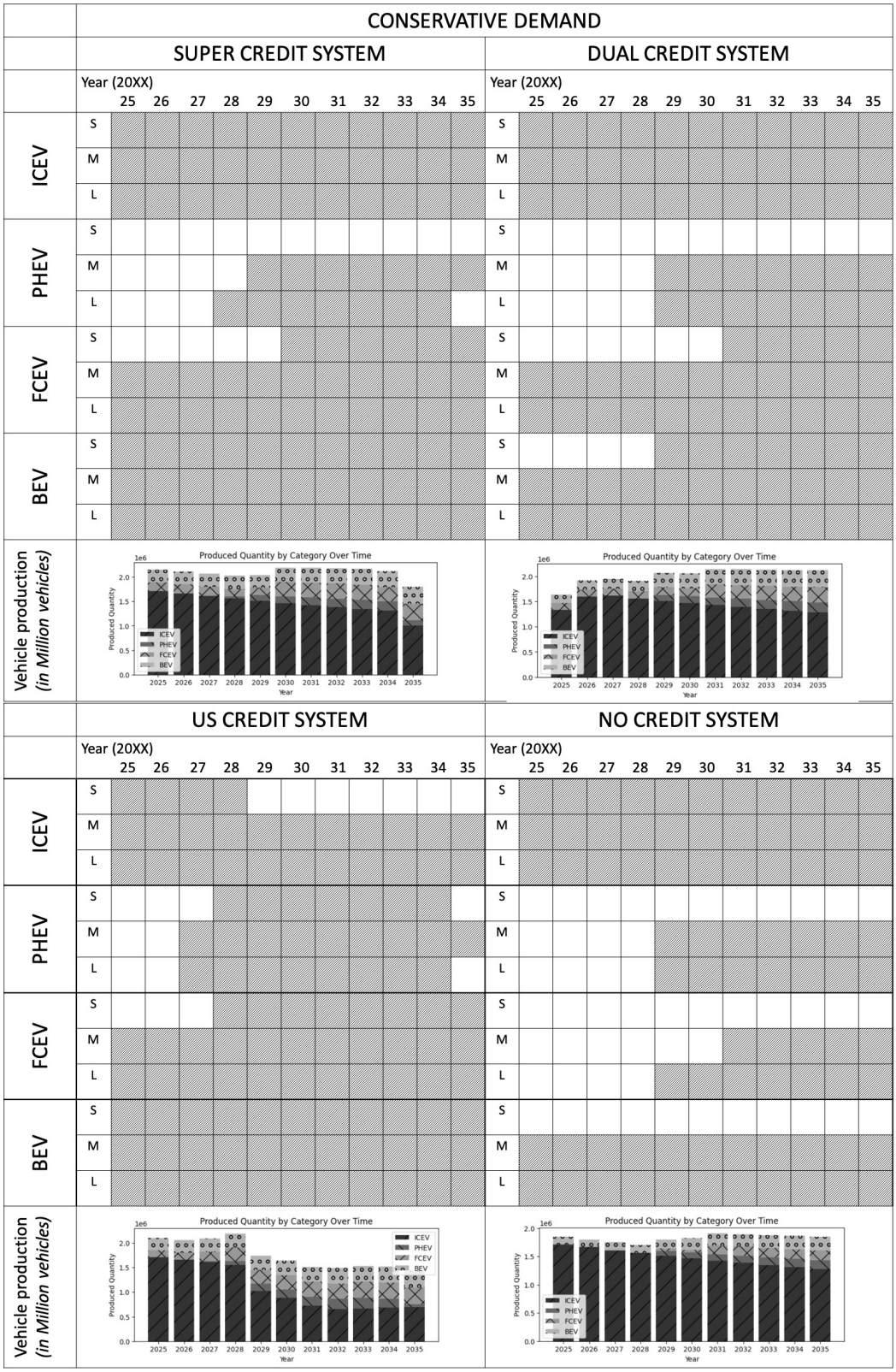
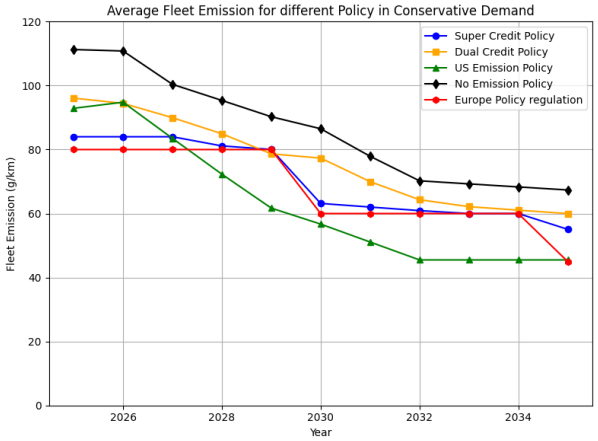
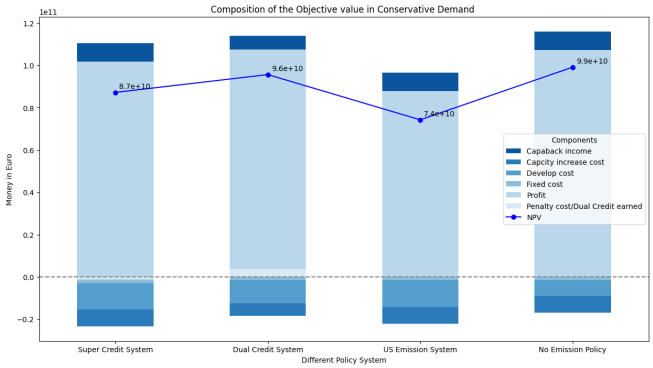


Figure 3: Optimal Portfolio for Different Policy in Conservative Demand

for all three policy scenarios, manufacturers would initialize a lower quantity of low-emission vehicle types, prioritizing those with larger profit margins to balance penalties and revenues. For example, small-sized PHEV vehicles would not be initialized in the US emission system. Similarly, the small-sized FCEV project would not start in the Super Credit system



(a) Fleet Emission



(b) NPV Structure

Figure 4: Policy Comparison in Conservative Demand

and would have fewer years in the Dual Credit System and US emission system. The same trend also applies to the small-sized BEV vehicle projects, with fewer projects started after 2029 in the Super Credit System due to a larger demand for other, more profitable low or zero-emission vehicle projects that work to balance the CO_2 emissions. Comparing the three policies, the Dual Credit policy is much less sensitive to the demand scenario for vehicle project initialization because of its credit system policy. The initialized vehicle projects in the optimal project portfolios for the innovate demand are similar to the projects for the conservative demand since there is no threshold but monetary incentives for manufacturers. For the Super Credit System and US Emission System, producers choose the vehicle type with the lowest emissions and the highest profit margin to balance the extra emissions from ICEV types and avoid penalty costs. Once the threshold is met, there is no incentive for manufacturers to produce additional low-emission vehicles. However, for the Dual Credit System, it is always profitable to produce more low-emission vehicles because manufacturers can earn more money for the extra credits earned.

The production quantity graph in Figure 5 also shows a similar trend. In the Dual Credit System, the total production quantity of FCEV and BEV vehicles consistently increases as the production of ICEV types decreases. However, for the Super Credit System and US Emission System, the total production quantity is lower due to reduced demand for ICEV types of vehicles. Additionally, in the innovative demand scenario, the composition of production quantities for different powertrain types does not change significantly. However, in the Dual Credit Policy scenario, the production quantity of zero-emission vehicles increases to earn more money through credits.

Fleet Emission and NPV Analysis

In the Innovative demand scenario, the line plot and bar plot were used to analyze fleet emissions and NPV values, as shown in Figure 6. The black line represents the baseline model with no emission policy. After 2031, due to increased market demand for low emission vehicles, fleet emissions naturally fall below the EU regulated threshold. For the Dual Credit Policy system (yellow line), fleet emissions are much lower, reaching their lowest point after 2032. In the Super Credit System (blue line), emissions follow the EU regulated threshold until 2031, after which they drop further due to market demand. The US Emission System (green line) shows a similar trend, but after 2032, the emission threshold is lower than the EU regulated threshold.

In the NPV structure graph, it can be observed that in the Innovative demand scenario, compared with the conservative demand situation, the differences in objective values between different policies are smaller. The Dual Credit Policy has a higher NPV value, about 4% more compared to the baseline model, due to the extra credits earned. For the Super Credit Policy and US Credit Policy, the objective function values are similar, both about 6% lower. The major reduction occurs before 2031, as after this year, the market itself becomes more favorable toward low emission vehicles, and the regulations have less or no effect on restricting manufacturers from producing more low emission vehicle types.

6.2. Sensitivity Analysis

In the context of sensitivity analysis for the policy factors, this paper selects up to three key factors for each emission policy. These factors are chosen based on their presumed significance on the policy outcomes and their potential for being readily adjusted by governmental authorities. For each factor analyzed, this study employs stacked bar charts to compare the Net Present Value (NPV) structure and the production quantities of different vehicle types. Additionally, it includes objective values and the percentage of low emission vehicles in the total production as represented in the line on the bar chart. The parameters are categorized into three levels with

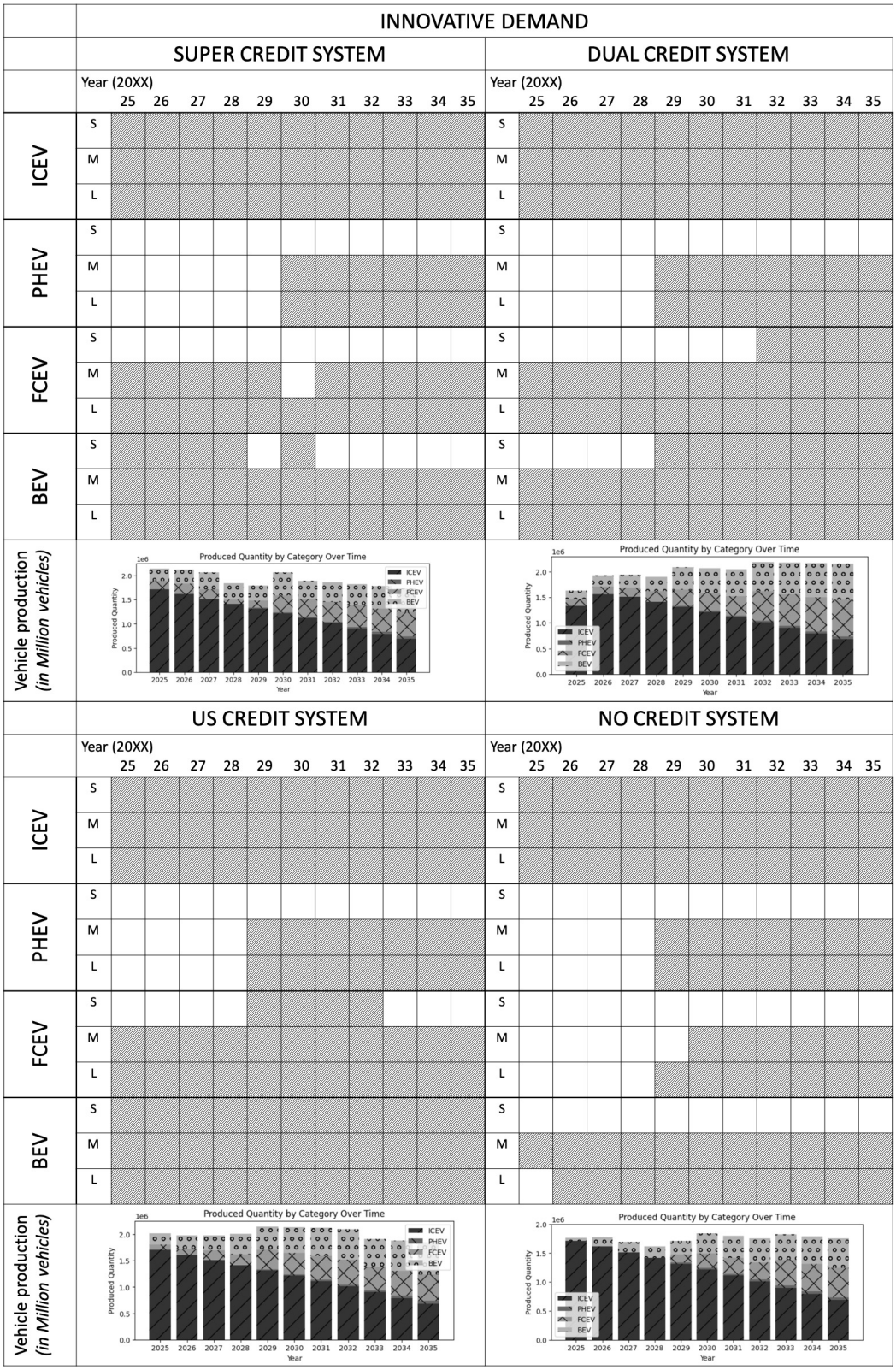
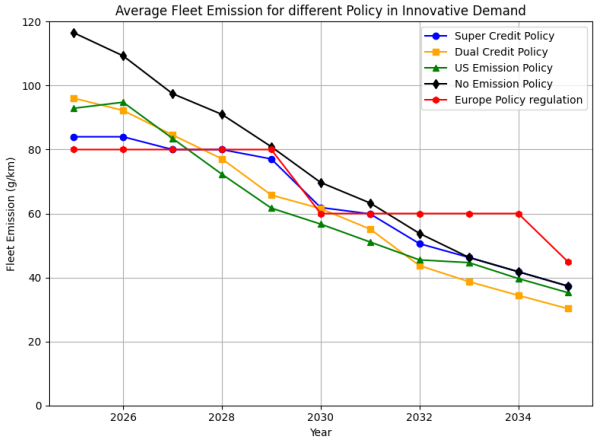
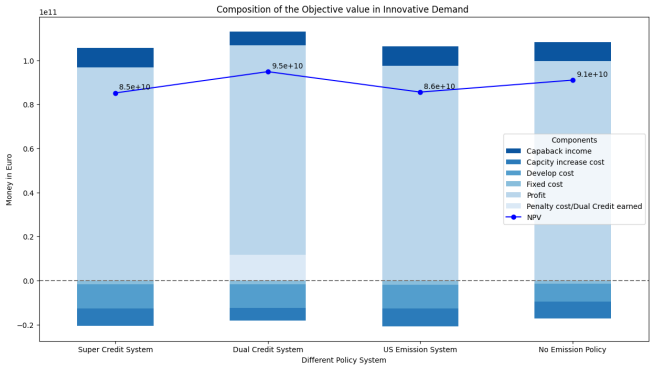


Figure 5: Optimal Portfolio for Different Policy in Innovative Demand

same intervals: high, basic, and low. The rationale behind this categorization is twofold: first, to reduce the experimentation process time, as each instance typically takes around five minutes to yield results, and second, the factors would not influence the trends. Further details regarding the parameter adjustment range can be found in Section 5.3.



(a) Fleet Emission



(b) NPV Structure

Figure 6: Policy Comparison in Innovative Demand

6.2.1. Super Credit System

Under the Super Credit System, this analysis focuses on three crucial factors: the regulated percentage of low emission vehicles, the maximum allowable relaxation percentage, and the PHEV multiplier. The regulated low emission vehicle percentage represents the minimum proportion of low emission vehicles (those emitting less than 50 g/km of CO2) that must be met before relaxation of the emission threshold is permitted. The maximum allowed relaxation percentage sets the upper limit for threshold relaxation. Finally, the PHEV multiplier determines whether PHEV-type vehicles receive a multiplier effect, meaning that when their CO2 emissions reach 50 g/km, they are counted as approximately 0.3 of a production unit instead of 0. The mathematical formulation for these factors can be found in equations 19.

Regulated EV Percentage

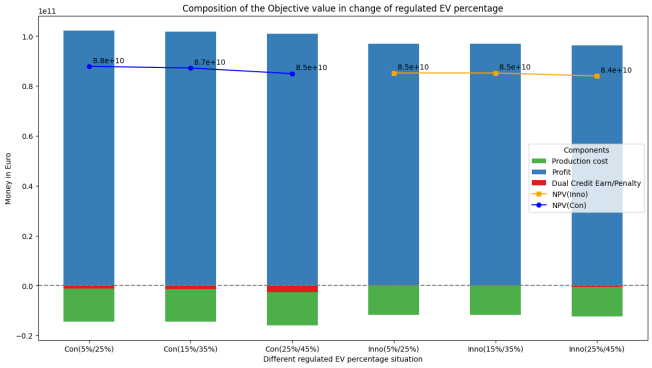
In the objective value diagram, I observe that as the required Regulated Low Emission Vehicle Percentage increases, the objective value is slightly affected, resulting in a decrease in the final NPV value. This trend is similar for both conservative and innovative demand scenarios, as shown in the Objective Value graph in Figure 7. However, the differences are smaller in the innovative demand scenario. The primary factor driving this cost difference is the penalty cost. As the required percentage increases, it becomes more challenging for manufacturers to achieve the goal percentage needed to benefit from the Super Credit Policy. This results in a higher penalty cost, which negatively impacts the NPV value.

In the production portfolio for different vehicle types, I observe a significant difference in the percentage of low emission vehicles produced as the threshold percentage is adjusted. When the threshold percentage is increased by about 10%, the percentage of low emission vehicles produced increases from 32.3% to 34.2%, representing a 2% increase. This change is mainly driven by the increase in the production volume of PHEV vehicles. On the other hand, when the threshold percentage is decreased, there is a slight reduction in the production of low emission vehicles, but this reduction is only about 0.4%, which is relatively small compared to the impact of increasing the threshold. Furthermore, in the innovative demand scenario, the production portfolio shows less variation and remains at a level of about 40%. This percentage level is higher than the percentage assumed for the conservative demand scenario and indicates that in the innovative demand scenario, a larger proportion of low emission vehicles is produced regardless of the threshold percentage.

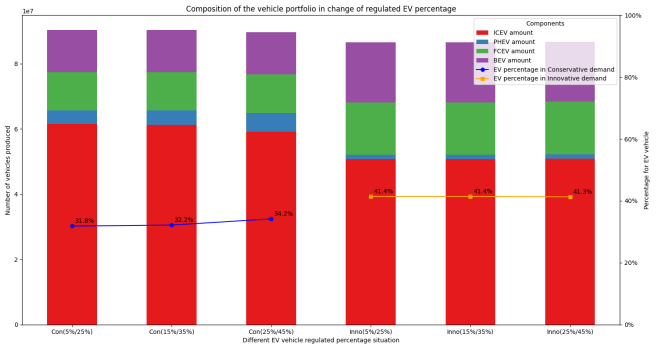
Maximum Allowed Relaxed Threshold

When I analyze the change in the maximum relaxed threshold allowed in the Super Credit System, I observe that as the allowed percentage increases, the objective value also increases. The primary difference is most noticeable in the penalty costs. This trend is more evident in the conservative demand scenario, as in the innovative demand scenario, manufacturers have a stronger incentive to produce low emission vehicles. In conclusion, the relaxation of Super Credit thresholds has a smaller impact on the production portfolio in the innovative demand scenario.

When examining the percentage of low emission vehicles produced in both innovative and conservative demand scenarios, I observe a stable trend with some slight differences. At allowed percentages of about 1% and 9%, the percentages are similar, likely due to changes in PHEV production. However, at higher or lower allowed percentages, manufacturers tend to produce more PHEV vehicles to increase their profits in conservative demand scenarios, while they produce fewer PHEV vehicles in innovative demand scenarios. The percentage of low emission vehicles is more stable in innovative demand, with only a 0.2% change compared to about 1% in conservative demand.

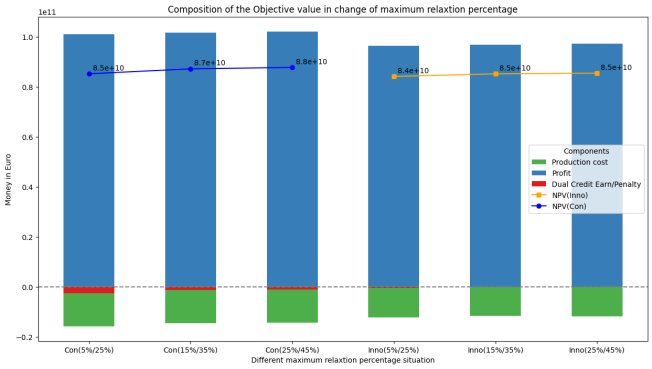


(a) Composition of the Objective Value

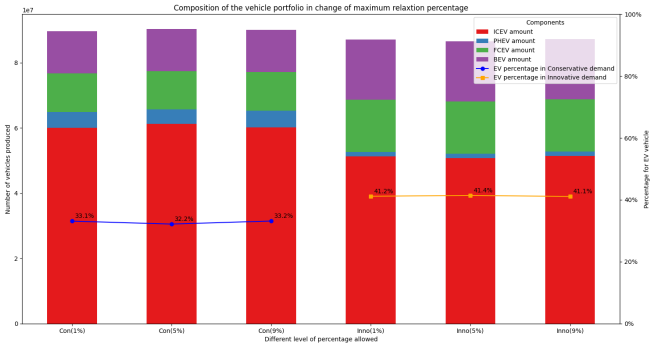


(b) Composition of the Vehicle Portfolio

Figure 7: Analysis for Change of Regulated EV Percentage



(a) Composition of the Objective Value



(b) Composition of the Vehicle Portfolio

Figure 8: Analysis for Change of Maximum Relaxation Percentage

PHEV Multiplier

In the implementation of the super credit policy, the introduction of the PHEV multiplier was intended to incentivize manufacturers to produce more PHEV vehicles. However, the results of the sensitivity analysis suggest that the PHEV multiplier may not significantly influence the behavior of manufacturers, possibly due to the relatively low market demand for PHEV vehicle types. The analysis shows that there are no significant differences in both the NPV graph and the vehicle portfolio graph. The composition and absolute values remain largely unchanged, with the conservative demand scenario consistently having a higher NPV objective value and about 10% fewer low emission vehicles produced.

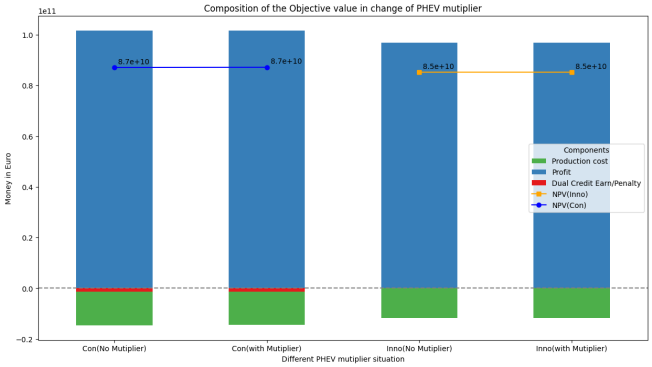
6.2.2. Dual Credit System

In the Dual Credit System, calculating dual credits involves complexity, and this paper simplifies certain factors by using average values. The system comprises three main parts, with one crucial factor selected from each part. These parts encompass the standard fuel consumption in calculating the CAFC score for traditional ICEVs, the credit exchange price, and the NEV weight factor in calculating the NEV score for low-emission vehicles (PHEV, FCEV, BEV). To analyze the impact of these factors, adjustments of approximately 50% compared to the current assumptions were made.

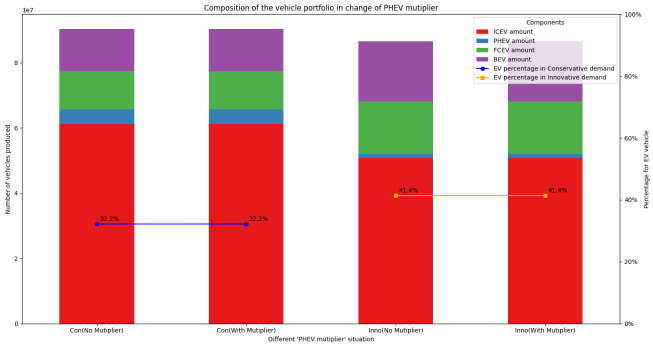
Standard Fuel Consumption (CAFC)

The change in the standard fuel consumption criteria has a notable impact on the manufacturer's objective value. With stricter restrictions, the objective value decreases because the manufacturer's ability to earn dual credits as extra profit diminishes. Conversely, as the standard fuel consumption index becomes more relaxed, the objective value increases. Specifically, a 50% decrease in the index leads to a 27% decrease in the objective value in a conservative demand scenario and an 18% decrease in an innovative demand scenario. Conversely, when the index becomes more relaxed, the objective value increases by approximately 16% in both demand scenarios.

Regarding the percentage of different vehicle types produced, in a conservative demand scenario, a stricter Standard Fuel Consumption index leads to a significant increase in the percentage of low emission vehicles produced, approximately 6%. Conversely, a more relaxed index results in a minor decrease in the percentage of low emission vehicles produced, about 0.4%. In an innovative demand scenario,

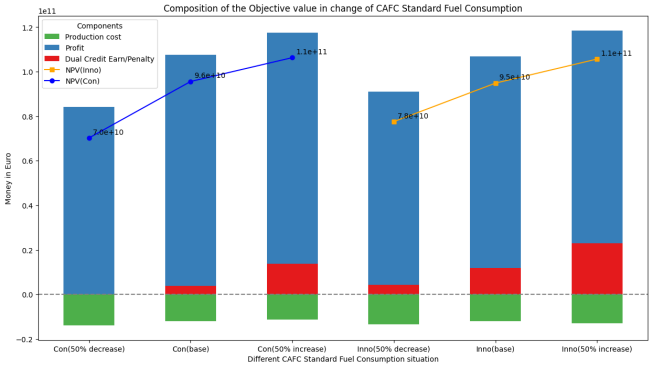


(a) Composition of the Objective Value

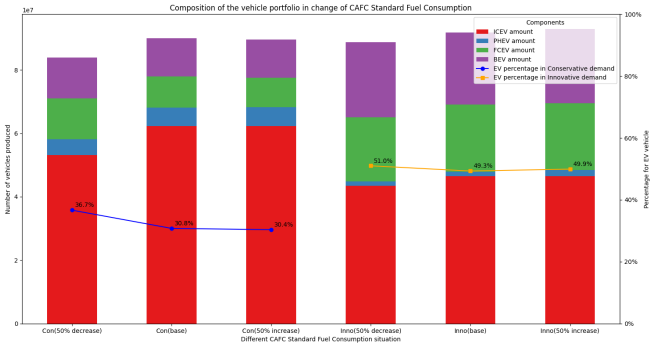


(b) Composition of the Vehicle Portfolio

Figure 9: Analysis for Change of PHEV Mutiplier



(a) Composition of the Objective Value



(b) Composition of the Vehicle Portfolio

Figure 10: Analysis for Change of CAFC Standard Fuel Consumption

a stricter index leads to a smaller increase of around 1% in the production of low emission vehicles, compared to a conservative demand scenario. A more relaxed index in the innovative demand scenario also results in a modest increase of approximately 0.6% in the percentage of low emission vehicles produced, along with an increase in total production volume to earn more dual credit.

Exchange Price

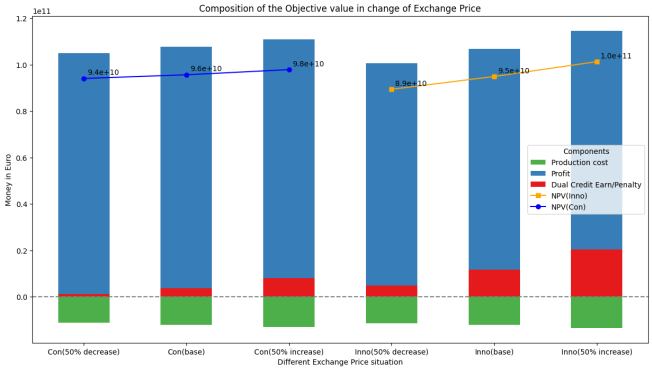
The change in the exchange price for the dual credit policy affects both the objective value and the production portfolio. When the exchange price increases, the objective value for the manufacturer also increases, but the extent of the increase is smaller compared to changes in the Standard Fuel Consumption index in CAFC credit calculation. This trend and value of increment are similar in both conservative and innovative demand scenarios, with about a 50% increase in the price resulting in about a 2% to 5% increase in NPV.

Regarding the production portfolio, as the price increases, manufacturers tend to produce more low emission vehicles to earn the dual credit value. The demand scenario does not significantly affect the trend, and the percentage increase tends to follow a logarithmic pattern rather than a linear one. With a larger price, there is a lower increase rate in the production of low emission vehicles.

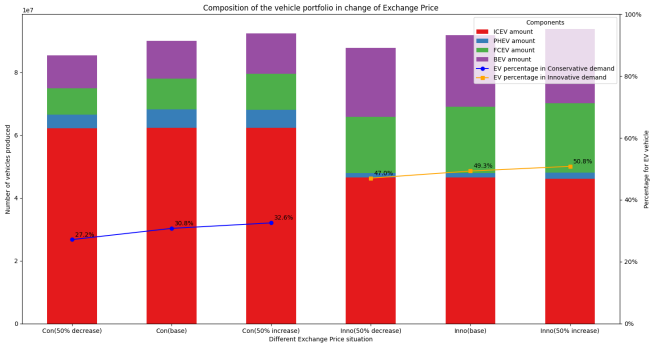
NEV Weighted Factor (NEV)

The NEV weight factor index is a critical factor in determining the NEV score in the dual credit policy system. This factor determines how much low emission vehicles are counted in calculating the NEV score. A higher NEV weight factor index score gives low emission vehicles a higher score in the NEV score calculation, which can help manufacturers earn more value. As the NEV weight factor index increases, the objective value also increases. This trend is consistent in both demand scenarios and increases linearly by about 4.

Regarding the percentage of the production portfolio, an increase in the NEV weight factor index motivates vehicle manufacturers to produce more low emission vehicles in both demand scenarios. The increase is also linear, with a slightly lower rate of increase in the innovative demand scenario, but it results in about a 1% increase in the absolute value in both scenarios. The increase in low emission vehicle production is more prominent for FCEV vehicles since FCEV has the highest NEV weight factor compared to other vehicle types.

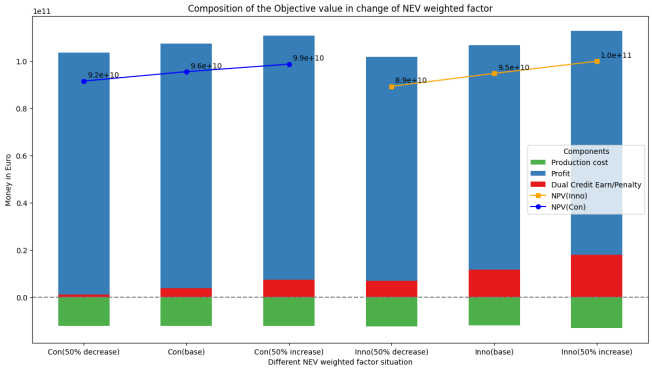


(a) Composition of the Objective Value

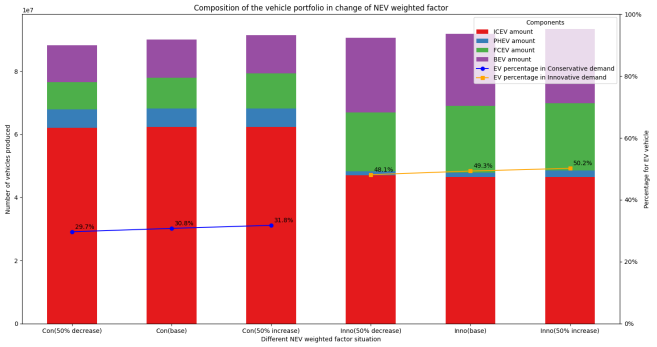


(b) Composition of the Vehicle Portfolio

Figure 11: Analysis for Change of Exchange Price



(a) Composition of the Objective Value



(b) Composition of the Vehicle Portfolio

Figure 12: Analysis for Change of NEV Weighted Factor

6.2.3. US Credit System

The US Credit System is a less complicated emission policy system that only includes a threshold for the average fleet emission volume in each year. However, this threshold is crucial in determining the effectiveness of this policy. A sensitivity analysis is performed on this threshold value to understand its impact on the policy's outcomes.

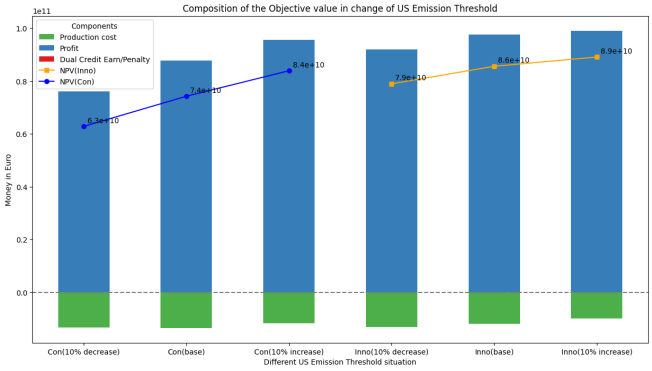
Average Fleet Emission Threshold

Figure 13 shows that the threshold for the average fleet emission has an impact on the objective value as well as the vehicle production portfolio. As the threshold decreases and becomes more stringent, the objective value decreases, with a larger decrease in the conservative demand scenario (about 14%) due to more restrictions on the production of ICEV vehicle types. In the innovative demand scenario, the decrease is about 8%. However, as the threshold is relaxed, the objective increases, to a greater extent in the conservative demand scenario due to the more relaxed restrictions. In the innovative demand scenario, the net objective value does not increase as much because market demand already motivates manufacturers to produce more low-emission vehicles.

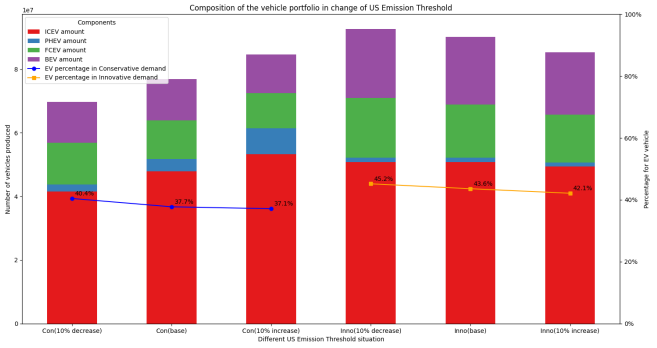
Regarding the vehicle production portfolio, allowing fewer fleet emissions leads manufacturers to produce a larger number of low-emission vehicles, a trend that applies to both conservative and innovative demand scenarios. It is also observed that in the conservative demand scenario, a more relaxed threshold encourages manufacturers to produce more ICEV and PHEV vehicles, resulting in a slower decrease in the total percentage of production for low-emission vehicles.

6.3. Factors Analysis

Factorial analysis was conducted to assess how different factors affect the final outcome of the policy. The results were analyzed with a focus on two key indicators: NPV value and EV percentage (the percentage of low-emission vehicle production). These indicators were chosen to evaluate how the factors impact both the manufacturer's profit and the government's environmental goals. The NPV value serves as a measure of how these factors affect the manufacturer's profitability. A higher NPV value is generally preferred by manufacturers, and the government also aims to ensure that manufacturers remain profitable to retain their presence in the country. The EV percentage is used to assess how well manufacturers comply with regulations related to producing environmentally friendly vehicles. A higher EV percentage indicates



(a) Composition of the Objective Value



(b) Composition of the Vehicle Portfolio

Figure 13: Analysis for Change of US Emission Threshold

that a greater number of low-emission vehicles are being produced, which contributes to lower overall CO2 fleet emissions. This aligns with the goals of emission policies to reduce environmental impact. By analyzing how these factors influence NPV and EV percentage, the car manufacturer can better understand the implications of different policy choices on both economic and environmental outcomes.

Three graphs have been generated to analyze the factors. The first one is a Pareto Chart, which is used to assess the absolute value of the importance for each factor and their cross-combinations. The factors are ordered from the most important to the least important, and a red vertical reference line is drawn to indicate which factor is statistically significant. The second plot is the Main Effect Plot, which checks how each factor affects the objective value output. Each factor is represented by a line with a different slope. A horizontal line indicates that the factor has no significant impact on the mean value of the chosen indicator. This allows for the visual analysis of the linear relationship between the factor and the indicator. The last graph is the Interaction Plot, which shows the relationships between different factors and the continuous indicator values. It is plotted in several blocks according to the factors' combinations. The main blocks to be analyzed are the three plots with a white background. By using this interaction plot, the effects of different factors can be compared under two different demand levels, which are considered uncertain factors that need to be estimated. Understanding the differences between different demand levels is essential for the car manufacturer to understand the policy effect. These three plots provide a thorough understanding of these factors and offer insights into the emission policy's details and outlook.

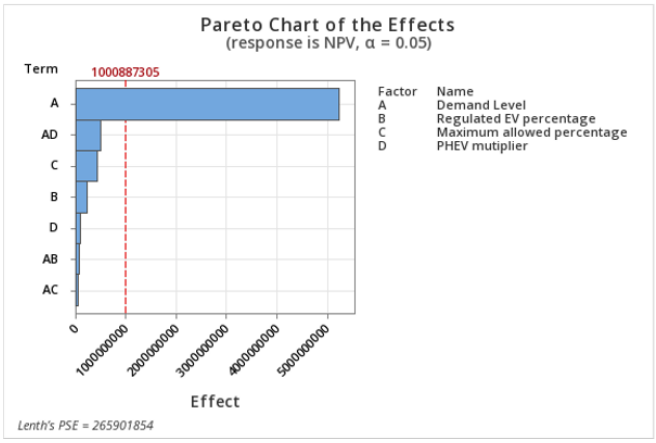
6.3.1. Super Credit System

In the Super Credit System, four factors have been selected, which are the same factors that were analyzed in the sensitivity analysis in section 6.2.1.

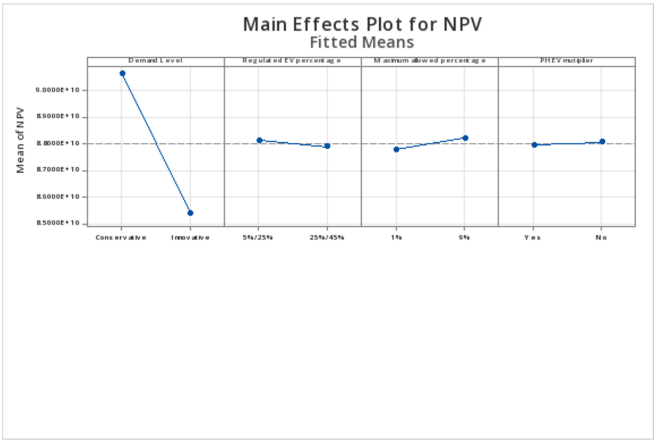
Net Present Value

Figure 14 and Figure 15 contain information related to the factorial analysis of the net present value (NPV) in the Super Credit policy. The Pareto graph in Figure 14 reveals that the demand level has the most significant impact on the net objective value of the vehicle manufacturer, with high statistical significance at the 0.05 level. For the other factors, namely Maximum percentage allowed, regulated EV percentage required, and the PHEV multiplier, their importance is not significantly different and lacks statistical significance at the current confidence level. Additionally, from the main effect chart in Figure 14, it is evident that under the Super Credit policy system, the objective value substantially increases with a conservative demand level, exhibiting a steep slope. Conversely, for the other three factors, the slope is nearly horizontal, indicating minimal differences between lower and higher levels. Lower regulated EV percentage thresholds result in slightly higher profits, while the trend for the maximum percentage allowed differs, resulting in lower objective values for lower percentages. The factor PHEV multiplier does not affect the objective value significantly and remains flat in the fitted means value comparison.

The interaction plot in Figure 15 reveals the interactive relationship between the factors. It is evident that, under different scenarios, the impact of the three factors related to the super credit system remains consistent. The slope lines for low and high values overlap, indicating that conservative demand consistently results in a higher objective value, regardless of variations in the factors. However, as the demand level changes, the effects of factor adjustments show some slight differences. Concerning the regulated EV percentage threshold, in the innovative demand scenario, there is a slightly more noticeable decrease in the objective level as the threshold percentage increases. The maximum percentage allowed does not exhibit differences between different demand scenarios. Yet, for the PHEV multiplier, the trend varies between the two demand scenarios. There is a positive relationship in the conservative demand scenario and a negative relationship in the innovative demand scenario. This suggests that



(a) Pareto Chart for Factor Analysis



(b) Main Effect Plot

Figure 14: Factor analysis in Super Credit System for NPV

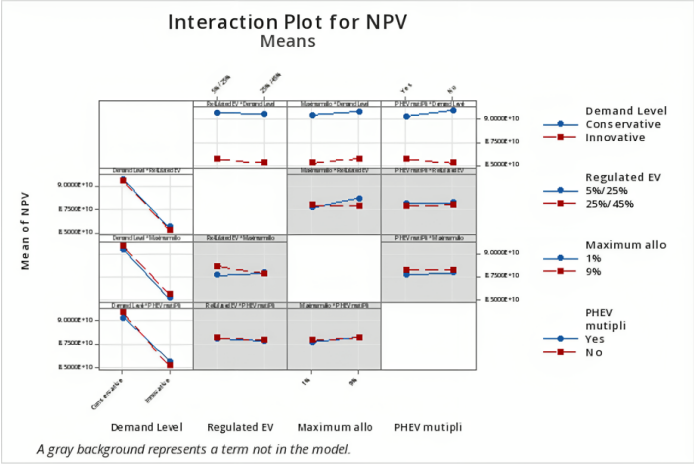


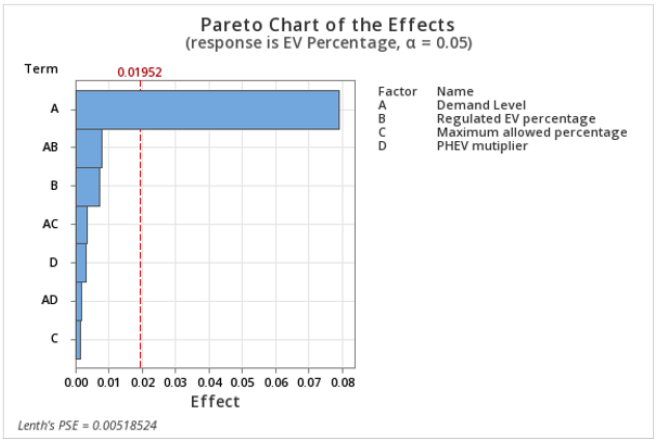
Figure 15: Interaction Plot of NPV in Super Credit System

the implementation of the PHEV multiplier would increase manufacturer profits in a conservative demand scenario but decrease profits in an innovative demand scenario. Although the slope remains flat in this experiment, it could increase as the policy maker assigns a higher value to the PHEV multiplier.

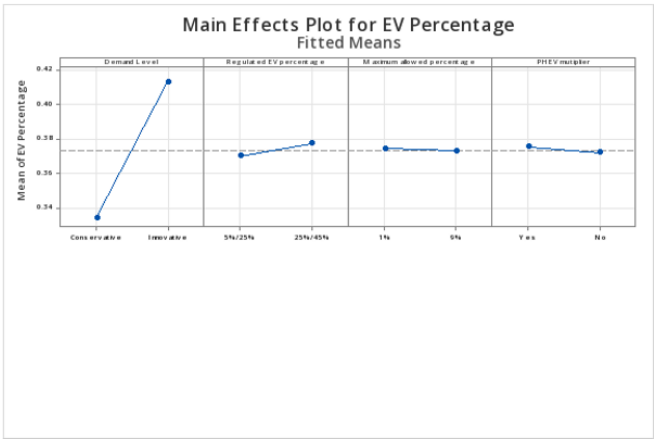
EV Percentage

Figure 16 and Figure 17 provide information regarding the factorial analysis for the percentage of low emission vehicles produced under the Super credit policy. The Pareto chart in Figure 16 indicates that the demand level is the only factor with statistical significance in altering the percentage of low emission vehicles (EVs) produced. The other factors are of similar importance. The main effect plot also illustrates that innovative demand results in a significantly higher percentage of electric vehicle production, represented by a steep slope. Conversely, for the other factors, such as the regulated EV percentage threshold, there is a slight increase in the production volume of electric vehicles. The maximum allowed percentage and PHEV multiplier yield horizontal lines, signifying no substantial differences in the level of electric vehicle production.

Furthermore, the interaction plot in Figure 17 suggests that the policy factor levels do not significantly affect the percentage of electric vehicles produced under different demand scenarios. Only the regulated EV percentage threshold exhibits a flatter slope in the innovative demand scenario, indicating that a lower regulated EV percentage threshold in conjunction with innovative demand results in slightly more electric vehicle production. Regarding the influence of the demand level on the policy factors, there are some minor differences. In the innovative demand scenario, changes in the three Super credit policy factors do not affect the percentage of electric vehicles produced, resulting in a horizontal line. However, in the conservative demand scenario, an increase in the regulated EV percentage threshold leads to higher electric vehicle production, while an increase in the maximum percentage allowed for relaxation and the implementation of



(a) Pareto Chart for Factor Analysis



(b) Main Effect Plot

Figure 16: Factor analysis in Super Credit System for EV percentage

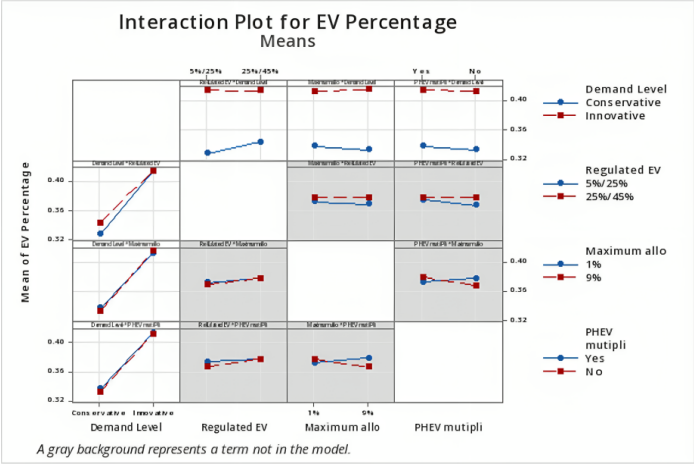


Figure 17: Interaction Plot of EV Percentage in Super Credit System

the PHEV multiplier results in a slight decrease in the percentage of electric vehicles produced. Nonetheless, the decrease in production volume in the latter cases is not significant at the current difference levels.

6.3.2. Dual Credit System

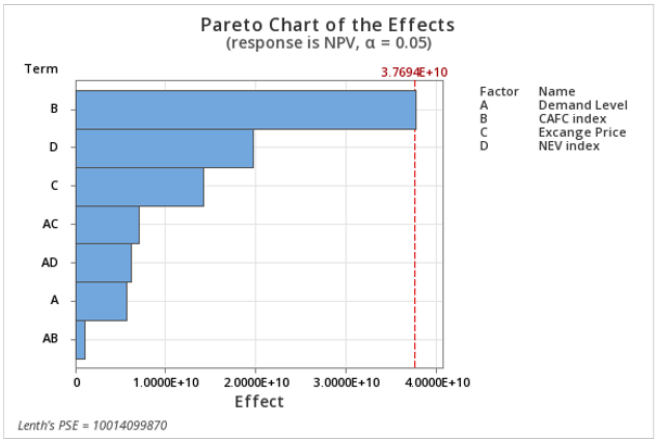
Under the dual credit system, four factors have been selected for analysis, and these factors align with those examined in the previous sensitivity analysis in Section 6.2.2.

Net Present Value

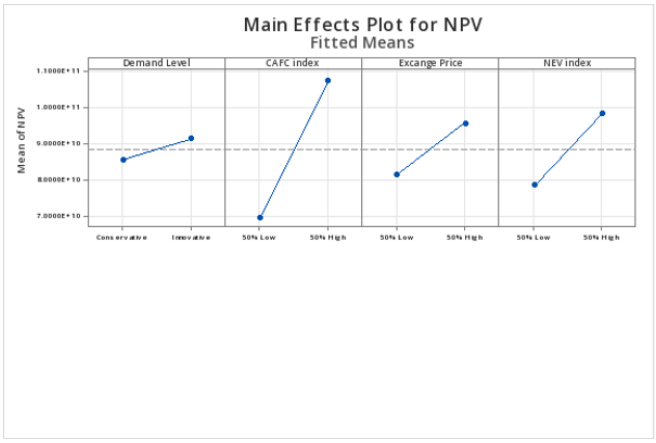
Figure 18 and Figure 19 provide information regarding the factorial analysis of the Net Present Value (NPV) under the Dual Credit policy. From the Pareto chart, it is evident that the most important factor affecting the manufacturer's NPV is the change in the CAFC index, which represents the standard fuel consumption value. This factor is statistically significant at a 95% confidence interval. The following factors, in descending order of importance, are the NEV index (NEV weighted factor), exchange price, and demand level. In the main effect plot, it can be observed that for the demand level and all three dual credit policy factors, there is a linear positive trend. As these factor values increase, the NPV also increases. However, unlike the Super Credit policy, the demand level exhibits only a minor difference, with innovative demand resulting in slightly higher NPV compared to conservative demand.

The interaction plot in Figure 19 is slightly more complex when compared to the Super Credit policy. It reveals the interplay between different factors at different levels and under varying demand scenarios.

For the CAFC index, it is evident that the demand level doesn't have a significant impact on the NPV value. Higher CAFC index values result in higher NPV values, regardless of the demand level. In the case of the exchange price and NEV index, the trend differs between the two demand scenarios. When the exchange price is high, both demand scenarios yield similar NPV values. However, as the exchange price decreases to a lower value, the innovative demand scenario leads to higher earnings for the manufacturer. Regarding



(a) Pareto Chart for Factor Analysis



(b) Main Effect Plot

Figure 18: Factor analysis in Dual Credit System for NPV

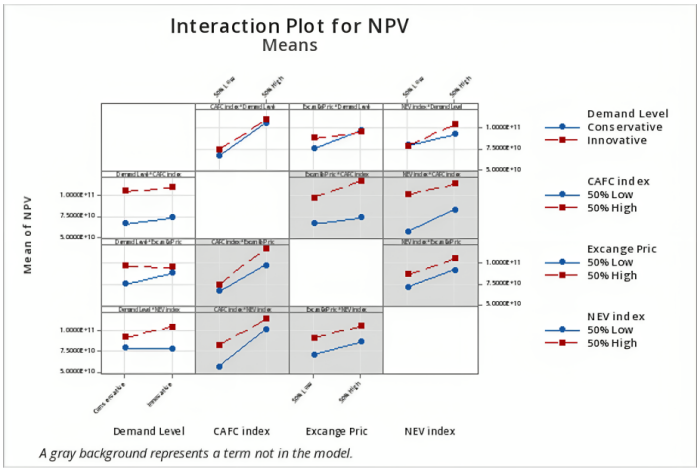


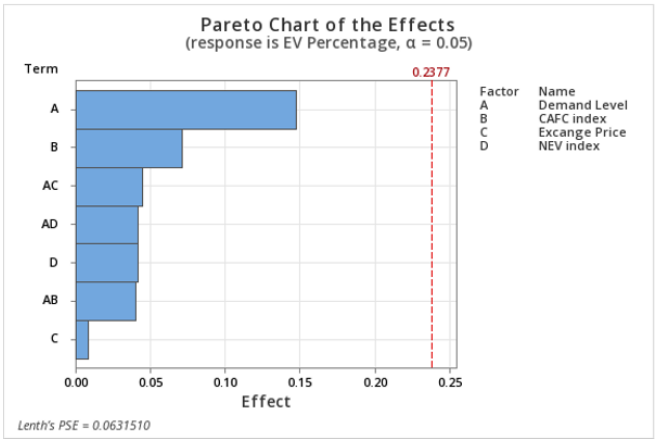
Figure 19: Interaction Plot of NPV in Dual Credit System

the NEV index, the trend is opposite to that of the exchange price. In the innovative demand scenario, a higher NEV index results in greater earnings. As the NEV index decreases to lower levels, the benefits diminish. The full matrix interaction plot also demonstrates the effects of changing policy factors under different demand levels. In a conservative demand scenario, the CAFC index exhibits the steepest positive slope, followed by the exchange price and NEV index, which have similar positive slopes. This suggests that increasing these factor values leads to higher earnings for the manufacturer. Conversely, in the innovative demand scenario, the CAFC index doesn't result in significant differences. However, the slope for the exchange price becomes flatter, indicating a lower increase in NPV value as the exchange price increases. The NEV index slope becomes steeper, signifying higher profits as the NEV index value increases.

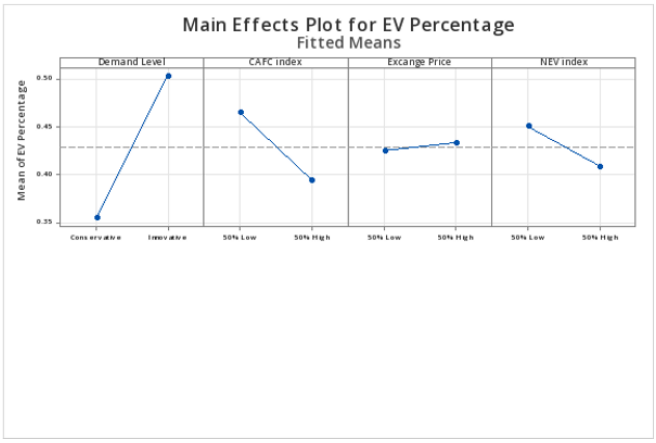
EV Percentage

Figure 20 and Figure 21 provide information on the factorial analysis for the percentage of low emission vehicles produced under the Dual credit policy. The Pareto Chart reveals that none of the factors have statistical significance in affecting the percentage of electric vehicles produced in the Dual emission policy system. The demand level is the factor with the most significant impact, but it does not reach statistical significance. From the main effect plot, it can be concluded that the innovative demand scenario results in a higher percentage of electric vehicle production. The CAFC index and the NEV index reduce the proportion of electric vehicles produced, with higher index values indicating greater tolerance for fuel-powered vehicles. However, the exchange price does not significantly affect the percentage of electric vehicles produced, as there is no observable difference as the price changes from low to high.

The interaction plot in Figure 21 shows that the values of the dual policy factors do not alter the trend of electric vehicle production in different demand levels. Innovative demand consistently leads to a higher percentage of electric vehicles in the production portfolio. The slope varies, with



(a) Pareto Chart for Factor Analysis



(b) Main Effect Plot

Figure 20: Factor Analysis in Dual Credit System for EV percentage

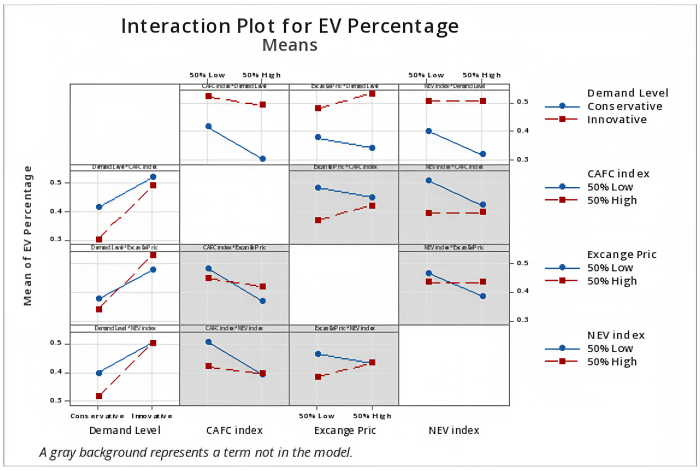
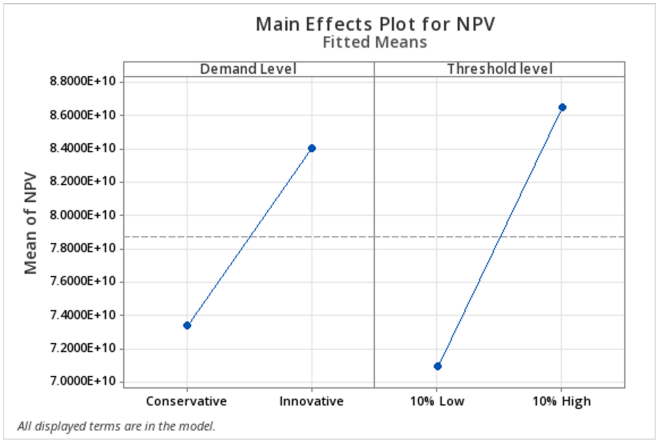


Figure 21: Interaction Plot of EV Percentage in Dual Credit System

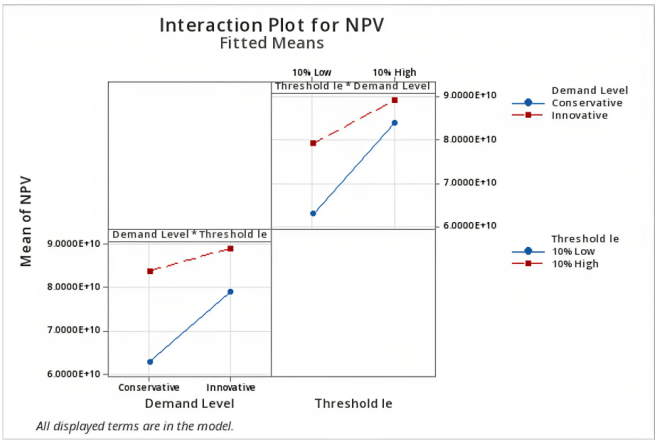
higher values of the three dual policy factors resulting in a greater increase in the percentage of electric vehicles in the production portfolio. Regarding the influence of changes in dual credit policy under different demand levels, in a conservative demand scenario, the percentage of electric vehicle production decreases as the factor values increase. The CAFC index exhibits the largest slope, followed by the NEV index and the exchange price. However, in the innovative demand scenario, the impact of changes in the CAFC index and NEV index is much lower, and for the exchange price, the percentage of electric vehicle production actually increases as the price increases. This may be due to increased market demand for low-emission vehicles, motivating manufacturers to produce more electric vehicles to earn from the dual credit score and achieve better profits.

6.3.3. US Credit System

The US credit system is simpler, with only one policy-related factor, which is the CO_2 emission threshold value. Figure 22 and Figure 23 depict the main effect and interaction plots for the net present value (NPV) and the electric vehicle (EV) percentage in the entire production portfolio. For the net present value, in the innovative demand scenario and with a higher level of the emission threshold, the manufacturer achieves a higher NPV. The slope is similar, indicating that the difference in threshold value has a consistent effect on the NPV regardless of the demand level. The interaction plot shows that in the conservative demand case, the NPV is more sensitive to changes in the threshold value, but the innovative demand scenario consistently results in a higher NPV. Regarding the electric vehicle percentage in the production portfolio, the innovative demand scenario yields a higher electric vehicle percentage. However, a higher threshold value leads to a lower proportion of electric vehicles produced. The effect of threshold changes on the electric vehicle production percentage is consistent in both demand scenarios, with innovative demand consistently resulting in a higher percentage of electric vehicles produced, albeit with some absolute value differences.

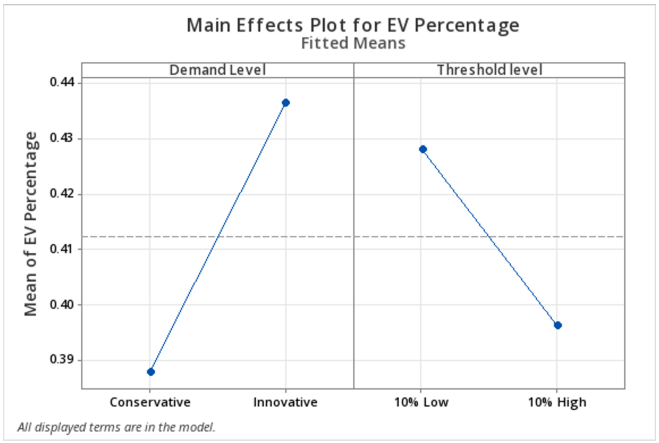


(a) Main Effect Plot

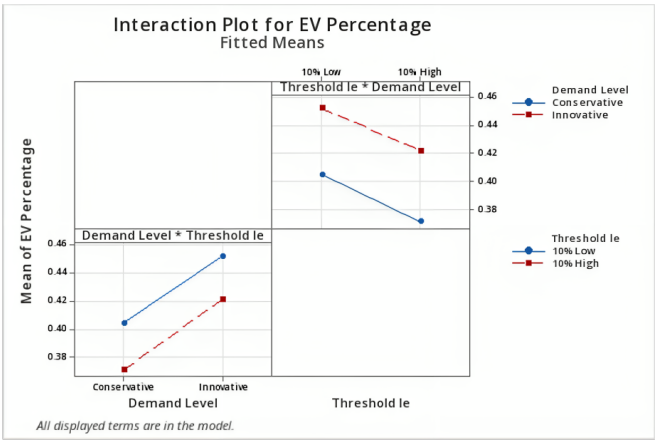


(b) Interaction Plot

Figure 22: Factor analysis in US Credit System for NPV value



(a) Main Effect Plot



(b) Interaction Plot

Figure 23: Factor analysis in US Credit System for EV percentage

7. Conclusion and Outlook

This article employs a mathematical model, primarily a mixed-integer linear model, to characterize car manufacturers' production portfolios under different emission policies. The aim is to gain a more quantitative understanding of various policy systems and their effectiveness. By integrating actual datasets with different demand scenarios and policy parameters, the study simulates the impact of these policies, assuming that manufacturers strive to maximize their profits.

The results indicate that all three policy systems in Europe, China, and the United States contribute to increased production of low-emission vehicles compared to the base model with no policy in place. Regarding the initialization of low-emission vehicles, the US emission policy leads to the most significant increase in the number of vehicle initiations, while the Super Credit Policy in Europe and the Dual Credit Policy perform similarly in this regard. However, from a financial perspective, the Dual Credit Policy performs the best in preserving the car manufacturer's profit, while the US emission policy has the most detrimental effect on the manufacturer's profit. For the average fleet emissions, both the Super Credit Policy and the US emission policy effectively track the trend set by the regulated emission threshold. However, the Dual Credit Policy does not have a fixed threshold but rather follows emissions according to market trends. As the market increasingly favors low-emission vehicles, resulting in a shift towards an innovative demand scenario where consumers prefer such vehicles, the fleet emissions are lower in response to this trend.

Furthermore, my experiments show that, it is evident that for the Super Credit Policy and the US Emission Policy, the demand level has a significant impact on the profit of the car manufacturer and the percentage of low-emission vehicles produced. In contrast, the Dual Credit Policy exhibits notable differences, where the demand level appears to influence primarily the percentage of low-emission vehicles produced but not the objective value.

Speak to the specific parameter factors, the Super Credit

System does not exert a substantial influence on either of the indicators, namely NPV (Net Present Value) and EV (Electric Vehicle) percentage. However, for the Dual Credit Policy and the US Emission Policy, these factors exhibit a relatively higher effect on both indicators, with the CAFC index being the most significant factor in the Dual Credit Policy. Notably, the market exchange price for the Dual Credit only appears to impact the net present value and not the percentage of low-emission vehicle production. In the context of the US emission policy, the only factor, which is the carbon fleet emission threshold, plays a significant role in determining the manufacturer's profit and the policy's effectiveness.

There are several limitations to this study. Firstly, regarding the model solving method, the Super Credit System policy is not formulated as a linear model and cannot be readily transformed into a linear form for optimization. Instead, a heuristic method was employed to solve this model in two steps. While this approach may not guarantee an optimal solution, however, my experiments showed that the heuristic gives good quality solutions that resulting deviations are unlikely to have a significant impact.

Secondly, in this study, market demand is based on simplified assumptions and is not specific to individual vehicle models. The total demand amount is assumed to be constant for each year in the planning period, and vehicle sales quantities may vary due to consumer preferences. Additionally, the vehicle type segmentation used in this study is relatively broad, categorizing vehicles based on powertrain technology, size, year, and power range. In reality, there is a much larger variety of vehicle models. To address this limitation, a more accurate demand prediction model could be integrated into the analysis to enhance result accuracy.

Thirdly, some of the policy-specific parameters are based on assumptions. For instance, in the Dual Credit Policy system, certain parameters such as fuel consumption standard in CAFC score calculation or the weighted factor for actual NEV credit calculation are calculated based on detailed descriptions of specific vehicle models and are not included in the dataset due to privacy concerns. Obtaining a more comprehensive dataset may require closer collaboration with vehicle manufacturers.

At last, in the real market scenario, the cooperation between the different vehicle manufacturer should also been considered. In this study, it is assumed that only one vehicle manufacturer exists and there would be any market trading activity occurs. These activities is hard to be modelled but if necessary could be added to this model for a more accurate result.

References

Ajanovic, A., & Haas, R. (2017). The impact of energy policies in scenarios on GHG emission reduction in passenger car mobility in the EU-15. *Renewable and Sustainable Energy Reviews*, *68*, 1088–1096. https://doi.org/10.1016/j.rser.2016.02.013

An, F., Earley, R., & Green-Weiskel, L. (2011). Global overview on fuel efficiency and motor vehicle emission. http://www.icet.org.cn/english/adminis/uploadfile/20113271414317751.pdf

Bergek, A., & Berggren, C. (2014). The impact of environmental policy instruments on innovation: A review of energy and automotive industry studies. *Ecological Economics*, *106*, 112–123. https://doi.org/10.1016/j.ecolecon.2014.07.016

Chen, Z., & He, H. (2022, February). Policy Brief - California-China Climate Institute [Accessed: 2023]. https://ccci.berkeley.edu/sites/default/files/China_Dual_Credit_Policy_Brief_Feb2022.pdf

Costa, C., Barbosa, J., Castro, H., Gonçalves, R., & Lanceros-Méndez, S. (2021). Electric vehicles: To what extent are environmentally friendly and cost effective? – comparative study by european countries. *Renewable and Sustainable Energy Reviews*, *151*, 111548. https://doi.org/10.1016/j.rser.2021.111548

Cuenca, R. M., Gaines, L. L., & Vyas, A. D. (2000). *Evaluation of electric vehicle production and operating costs* (tech. rep.). https://doi.org/10.2172/764206

Dominioni, G., & Faure, M. (2022). Environmental Policy in Good and Bad Times: The Countercyclical Effects of Carbon Taxes and Cap-and-Trade. *Journal of Environmental Law*, *34*(2), 269–286. https://doi.org/10.1093/jel/eqac003

European Commission. (2023). CO2-Emissionsnormen für Personenkraftwagen und leichte Nutzfahrzeuge [Accessed: 2023]. https://climate.ec.europa.eu/eu-action/transport/road-transport-reducing-co2-emissions-vehicles/co2-emission-performance-standards-cars-and-vans_de

Hill, N., Skinner, I., Zazias, G., & Siskos, P. (2018, February). Assessing the impacts of selected options for regulating CO2 emissions [Transport & Environment Policy Research]. https://climate.ec.europa.eu/system/files/2018-03/ldv_post_2020_co2_en.pdf

Hu, Y., Wang, Z., & Li, X. (2020). Impact of policies on electric vehicle diffusion: An evolutionary game of small world network analysis. *Journal of Cleaner Production*, *265*, 121703. https://doi.org/10.1016/j.jclepro.2020.121703

Huang, S., Shen, H., Chen, Y., Chen, M.-a., & Li, X.-j. (2015). Global organic carbon emissions from primary sources from 1960 to 2009. *Atmospheric Environment*, *122*, 505–512. https://doi.org/10.1016/j.atmosenv.2015.10.017

International Council on Clean Transportation (ICCT). (2016). China CAFC and NEV Credits - International Council on Clean Transportation [Accessed: 2023]. https://theicct.org/sites/default/files/publications/China%20CAFC%20and%20NEV%20credits_policy-update_vF_18052018_updated.pdf

International Energy Agency (IEA). (2021). Fuel Economy in China – Analysis [Accessed: 2023]. https://www.iea.org/articles/fuel-economy-in-china

Kellner, F., Lienland, B., & Utz, S. (2021). A multi-criteria decision-making approach for assembling optimal powertrain technology portfolios in low GHG emission environments. *Journal of Industrial Ecology*, *25*(6), 1412–1429. https://doi.org/10.1111/jiec.13148

Kieckhäfer, K., Volling, T., & Spengler, T. S. (2012). Supporting Strategic Product Portfolio Planning by Market Simulation. In *Gabler Verlag, Wiesbaden* (pp. 123–147). https://doi.org/10.1007/978-3-8349-3722-3_6

Kieckhäfer, K., Walther, G., Axmann, J., & Spengler, T. (2009). Integrating agent-based simulation and system dynamics to support product strategy decisions in the automotive industry. *Proceedings of the 2009 Winter Simulation Conference (WSC)*, 1433–1443. https://doi.org/10.1109/WSC.2009.5429295

Lipman, T. E., & Delucchi, M. A. (2006). A retail and lifecycle cost analysis of hybrid electric vehicles. *Transportation Research Part D: Transport and Environment*, *11*(2), 115–132. https://doi.org/10.1016/j.trd.2005.10.002

Ma, H., Lou, G., Fan, T., Chan, H. K., & Chung, S. H. (2021). Conventional automotive supply chains under China's dual-credit policy: Fuel economy, production, and coordination. *Energy Policy*, *151*, 112166. https://doi.org/10.1016/j.enpol.2021.112166

Michalek, J. J., Papalambros, P. Y., & Skerlos, S. J. (2005). A Study of Fuel Efficiency and Emission Policy Impact on Optimal Vehicle Design Decisions. *Journal of Mechanical Design*, *126*(6), 1062–1070. https://doi.org/10.1115/1.1804195

Moran, D., Wood, R., Hertwich, E., Mattson, K., Rodriguez, J. F. D., Schanes, K., & Barrett, J. (2020). Quantifying the potential for consumer-

oriented policy to reduce European and foreign carbon emissions. *Climate Policy*, *20*(sup1), S28–S38. https://doi.org/10.1080/14693062.2018.1551186

Ou, S., Lin, Z., Qi, L., Li, J., He, X., & Przesmitzki, S. (2018). The dual-credit policy: Quantifying the policy impact on plug-in electric vehicle sales and industry profits in China. *Energy Policy*, *121*, 597–610. https://doi.org/10.1016/j.enpol.2018.06.017

Parikh, A., Shah, M., & Prajapati, M. (2023). Fuelling the sustainable future: a comparative analysis between battery electrical vehicles (BEV) and fuel cell electrical vehicles (FCEV). *Environmental Science and Pollution Research*, *30*(20), 57236–57252. https://doi.org/10.1007/s11356-023-26241-9

Register, F. (2023). 29184 Federal Register / Vol. 88, No. 87 / Friday, May 5, 2023 ... - GovInfo [Accessed: 2023]. https://www.govinfo.gov/content/pkg/FR-2023-05-05/pdf/2023-07974.pdf

Romejko, K., & Nakano, M. (2017). Portfolio analysis of alternative fuel vehicles considering technological advancement, energy security and policy [Cleaner production towards a sustainable transition]. *Journal of Cleaner Production*, *142*, 39–49. https://doi.org/10.1016/j.jclepro.2016.09.029

Thiel, C., Nijs, W., Simoes, S., Schmidt, J., van Zyl, A., & Schmid, E. (2016). The impact of the EU car CO2 regulation on the energy system and the role of electro-mobility to achieve transport decarbonisation. *Energy Policy*, *96*, 153–166. https://doi.org/10.1016/j.enpol.2016.05.043

Thies, C., Runde, C., Kieckhäfer, K., Wansart, J., & Spengler, T. (2022). Project portfolio planning under CO2 fleet emission restrictions in the automotive industry. *Journal of Industrial Ecology*. https://doi.org/10.1111/jiec.13228

Tietge, U., Díaz, S., Yang, Z., & Mock, P. (2017, November). From Laboratory to Road - International Council on Clean Transportation [Accessed: 2023]. https://theicct.org/wp-content/uploads/2021/06/Lab-to-road-intl_ICCT-white-paper_06112017_vF.pdf

TransportPolicy.net. (2018). China: Light-Duty: Fuel Consumption [Accessed: 2023]. https://www.transportpolicy.net/standard/china-light-duty-nev/

Vacheva, G., & Hinov, N. (2019). An overview of the state of art of fuel cells in electric vehicles. *2019 International Conference on Creative Business for Smart and Sustainable Growth (CREBUS)*, 1–4. https://doi.org/10.1109/CREBUS.2019.8840075

Wang, S., Zhao, F., Liu, Z., & Hao, H. (2018). Impacts of a super credit policy on electric vehicle penetration and compliance with China's corporate average fuel consumption regulation. *Energy*, *155*. https://doi.org/10.1016/j.energy.2018.05.042

Wang, Y., Zhao, F., Yuan, Y., Hao, H., & Liu, Z. (2018). Analysis of typical automakers' strategies for meeting the dual-credit regulations regarding CAFC and NEVs. *Automotive Innovation*, *1*(1), 15–23. https://doi.org/10.1007/s42154-018-0010-3

Yang, D.-X., Yang, L., Chen, X.-L., Wang, C., & Nie, P.-Y. (2023). Research on credit pricing mechanism in dual-credit policy: Is the government in charge or is the market in charge? *Environment, Development and Sustainability*, *25*(2), 1561–1581. https://doi.org/10.1007/s10668-022-02108-w

Yoro, S., & Al., E. (2020). *CO2 emission sources, greenhouse gases, and the global warming effect*. Woodhead Publishing. https://doi.org/10.1016/B978-0-12-819657-1.00001-3

Zhu, X., Ding, L., Zhu, H., & Guo, Y. (2022). The dual-credit policy model, a production strategy decision-making algorithm and application to Chinese automakers. *Neural Computing and Applications*. https://doi.org/10.1007/s00521-022-07703-6

Junior Management Science 10(3) (2025) 781-809

Junior Management Science

www.jums.academy
ISSN: 2942-1861



Numerical Studies for the Scheduling of Continuous Annealing Lines

Hagen Alexander Hönerloh

Leibniz University Hannover

Abstract

The continuous annealing of flat steel improves its properties for applications such as automotive manufacturing. Scheduling these processes on Parallel Heterogeneous Annealing Lines (PHALs) is complex due to diverse coil properties, incompatible process modes, and due date constraints. Introducing stringers to address incompatibilities between steel sheets raises costs, energy use, and CO2 emissions, highlighting the need for optimized scheduling. This thesis implements a mathematical model in Python using the Gurobi solver to optimize PHAL scheduling by minimizing stringer usage while meeting tardiness constraints. The model is extended to include coil-specific release dates and expanded to address trade-offs between stringer use, tardiness, and due date deviations, including earliness. A computational study evaluates the model under various scenarios, examining the effects of coil heterogeneity, urgency, process flexibility, and stringer processing times. Results show that optimized schedules reduce stringer use and delays, particularly under high process flexibility. These findings demonstrate the potential of optimization to improve efficiency and sustainability in steel production while guiding future research in dynamic scheduling approaches.

Keywords: continuous annealing lines; Gurobi solver; scheduling optimization; steel industry; stringer minimization

1. Introduction

1.1. Subject and motivation

Our modern economy thrives on digital transformation and its many facets. One key aspect is the computerization of processes using advanced digital technologies. Through computerization, companies can create and implement optimal schedules for their processes and thereby increase efficiency and productivity, as companies always sought to improve their decision-making through new scheduling and planning methods.[1] But the potential impact of computerization and digitalization could be far greater than anything before.

The steel industry is one of the many industries that stand to benefit significantly from this development. By leveraging digital technologies, manufacturers are able to minimize production time, while also maximizing the use of resources and thus their profits.[2] The impact on this highly competitive industry, which supplies us with materials needed for everyday appliances, railways, or even buildings, is astonishing.[3] One process in the steel industry, that can benefit greatly from digitalization, is the continuous annealing of flat steel, a method of processing steel to change its physical and mechanical properties.[4] In this process, coils of cold rolled flat steel are processed in furnaces with different annealing temperatures and transport speeds that make up the process mode of an individual coil. Different process modes lead to different mechanical and physical properties of the steel.[4] Hence, the mode is chosen according to the desired characteristics. Cold rolled flat steel is essential for building cars, and household appliances and has many more areas of application.[5] Scheduling the continuous annealing process on continuous annealing lines (CALs) is a difficult task due to the different properties of the flat steel, processing modes, and other aspects like due dates. If the properties or processing modes of two successive coils are too different, scrap

[1] Tang and Meng, 2021, p. 1
[2] Iannino et al., 2021, p. 620
[3] Zhao and Yang, 2016, p. 3417 and Terence, 2020
[4] Sarna, 2013
[5] Commodity-Inside, 2019

DOI: https://doi.org/10.5282/jums/v10i3pp781-809

© The Author(s) 2025. Published by Junior Management Science.
This is an Open Access article distributed under the terms of the CC-BY-4.0 (Attribution 4.0 International). Open Access funding provided by ZBW.

coils, so-called stringers, must be added between the coils to bridge these differences, resulting in additional material costs and a loss of efficiency.[6]

The impact of this loss of efficiency cannot be overstated. Not only does the manufacturer lose valuable production time, but he also has to waste energy, increasing the amount of CO_2 emitted per kg of steel. With a share of seven % of the world's annual CO_2 emissions, the steel industry is already one of the most energy-intensive industries worldwide.[7] In the meanwhile, in the European Union, the steel sector is facing an ever-increasing cost of energy, as well as an increase in $CO2$ price per ton of $CO2$ emitted, which some experts predict could reach 50% by the end of the decade.[8]

Hence, manufacturers should have a serious interest in optimizing their production schedules by minimizing the introduction of stringers and thus the costs caused by material and energy wastage, loss of efficiency, and CO_2 emissions, which can be achieved through the use of digital technologies.

1.2. Research question and structure of the work

This thesis is based on a paper by Wegel et al. (2024), in which they propose a mathematical model to optimize the scheduling of the continuous annealing process. The objective of this model is to minimize the introduction of stringers in a schedule, while also considering a tardiness constraint that limits the number of delays. It is designed for short-term planning and can therefore be used at the operational level of operations management.[9] So far, this model has only been implemented in the Julia programming language and used with a proprietary algorithm. Thus, the model has not yet been implemented in the popular Python programming language. This implementation forms the basis of this thesis.

During the course of this work, a numerical study will be conducted on this model and on its expansions created during the work. It will consist of different sections, which study the impact of certain parameters and scenarios on the scheduling of CALs. The results will be thoroughly analyzed and discussed to derive emerging trends and formulate managerial insights.

First, the CALs for flat steel will be explained in terms of their design and their scheduling, which will be followed by the current state of research. Afterwards, an explanation of the underlying problem and the mathematical model itself will be given. This will be continued by further extensions of the model with the aim of mimicking the real-world process. The numerical studies conducted will be the core of this thesis. Different trends that occur with increasing instance sizes will be explored first, followed by sections in which the impact of scenarios and the alteration of certain parameters on the scheduling process will be investigated. Furthermore, the base model will be compared with the extended model regarding its performance and solutions. The results of the computational study carried out on the Python implementation and the instances used are presented and discussed, leading to a summary of the work and an outlook for future research, further extensions, and managerial insights.

2. Continuous Annealing Lines for Flat Steel

2.1. Process description

The industrial continuous annealing process is made possible by CALs, which consist of several sections. One such CAL is depicted in figure 1. CALs are not standalone, they are part of a greater complex consisting of different areas dedicated to different processing methods, like the cold rolling and galvanization of steel.[10] This study will focus on the annealing of cold rolled steel, which is usually the bottleneck of steel processing.[11] In some cases, it may also be possible and advantageous to anneal hot rolled steel.[12] In comparison to hot rolling, cold rolling happens at far lower temperatures. But these processes are complementary and not substitutional, since the hot rolling process is an upstream process of cold rolling. While hot rolling rolls the steel to the desired width, cold rolling reduces its thickness by up to 90 %, increasing its strength and hardness but also severely diminishing its ductility and increasing its brittleness.[13] To improve these mechanical and physical properties, the steel strips are annealed in CALs. Through recrystallization of microstructures and other processes, the steel regains some of its lost characteristics, especially its ductility.[14]

The process starts at the entry section with the coils of cold rolled flat steel and the so-called pay-off reel, as depicted in figure 1. These coils stem from the upstream cold rolling process. The pay-off reels fixate the coils of flat steel with a mandrel and rotate to continuously unwind them.[16] Further downstream, a welding machine automatically welds the tail end to the head end of two consecutive coils together, thus providing continuous strip feeding to the succeeding sections.[17] The welding process itself is of the utmost importance since a weld break could result in a complete line shutdown.[18] Therefore, compatibility between consecutive coils regarding their thickness and width has to be ensured. If the two consecutive coils are incompatible with each other, a stringer has to be added between them to assure a continuous annealing process, resulting in a loss of efficiency and higher material and energy costs.[19] Stringers can be reused, but only for a limited number of times.[20]

[6] Besson, 1998, p. 29
[7] Joint-Research-Centre, 2022
[8] Twidale et al., 2021 and Krukowska, 2021
[9] Karakostas et al., 2020, p. 2 and Karakostas et al., 2019, p. 2
[10] Zhao and Yang, 2016, p. 3418
[11] Li et al., 2023, p. 1
[12] Steel-Warehouse, n.d.
[13] Zhao and Yang, 2016, p. 1
[14] Sarna, 2013
[15] Takurou et al. (2016), p. 113
[16] Jiyuan-Shenzhou-Industry, n.d.
[17] United-Enterprises, n.d.
[18] Williamson, n.d., p. 3
[19] Besson, 1998, p. 29
[20] Mujawar et al., 2004, p. 1

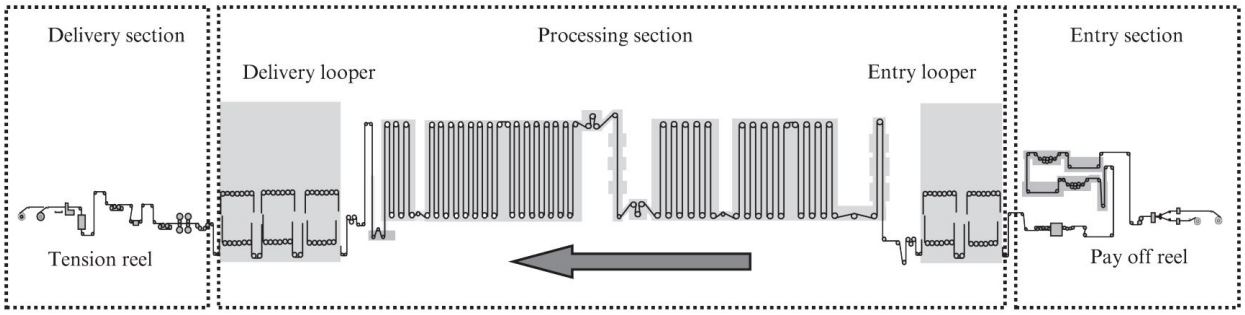


Figure 1: A Continuous Annealing Line as implemented in a Japanese factory.[15]

After degreasing by a degreasing unit, the steel strip enters the process section, where it first feeds into an entry looper.[21] This looper counteracts interruptions in the continuous annealing process, such as the welding process, and maintains a continuous strip speed throughout the annealing process. It achieves this by moving its rolls apart from each other, thus increasing the length of strip steel it can hold and the distance the steel has to traverse. The continuous strip then enters several furnaces, that can reach temperatures of up to 850 °C, while maintaining a strip speed of up to 800 m/min.[22] The steel strip cycles through the furnaces for several minutes, depending on the chosen processing mode. As previously described, a processing mode is a combination of annealing temperature and strip speed, which leads to certain steel properties.[23] Since the strip speed and furnace temperature can only be adjusted in a certain range from the preceding coil to the succeeding one, compatibility between their processing modes is necessary or else a stringer has to be introduced between them.[24] It should also be mentioned, that adjusting the furnace temperature costs a lot of energy and should therefore be minimized.[25]

After the heat treatment, the steel strip is cooled down in several steps, until it feeds into the delivery or exit looper.[26] This looper works like the entry looper and can therefore compensate for interruptions like maintenance and the cutting process, which follows downstream.[24] After the cutting, the tension reel recoils the steel strip into the previous coils.[27] Compared to the batch annealing process, in which the steel is processed as a coil, the continuous annealing process has a higher efficiency and productivity, while also delivering a more uniform product, regarding the physical and mechanical properties of the strip steel. Some of its disadvantages, however, are the large amounts of space and capital needed to construct it.[23]

[21] United-Enterprises, n.d.
[22] United-Enterprises, n.d.
[23] Sarna, 2013
[24] Besson, 1998, p. 29
[25] Zhao and Yang, 2016, p. 3417
[26] Sarna, 2013
[27] Jiyuan-Shenzhou-Industry, n.d.

2.2. Differentiation from literature

The topic of scheduling has been a research subject since the early 20th century and is nowadays one of the most researched fields in operations research, with several hundred papers published each year.[28] Scheduling in the steel industry particularly is one of the most difficult and complex problems, due to the complexity of the steel industry itself and therefore, there have been many attempts to optimize certain aspects of it.[29] This section will feature different approaches concerning CALs and continuous galvanizing lines, a process further downstream of the annealing process that has many similarities with CALs regarding its scheduling.[30]

Li et al. (2023) aimed to minimize earliness and tardiness costs, as well as setup costs, which occur through changes in annealing temperature, on one processing line.[31] The authors noted, that these objectives conflict with each other since minimizing the setup costs by constructing large batches of similar coils could increase the earliness and tardiness costs and vice versa. Stringers were not directly considered. To minimize the three objectives, they used an adaptive multi-objective differential evolutionary algorithm (MODE) based on deep reinforcement learning (DRL). Several other papers also include the use of MODEs for the optimization of CALs and for other processes as well, like for the hot rolling process.[32] The usage of MODEs can be highly effective, as can be seen in a study by Dong et al. (2021), in which they were able to find optimal schedules for the color-coating process for up to 400 coils with three different objectives. During the color coating process, protective or decorative coatings are applied to the steel.[33]

Zhao and Yang (2016) tested a discrete differential evolution algorithm (DDE), which uses discrete job permutations to find optimal solutions, to average the line capacity and minimize the changeover costs. These occur through annealing temperature changes in the furnace, and Zhao and Yang (2016) concluded that their algorithm performed well for up to 90 coils on parallel processing lines.[34] Even though the dif-

[28] Potts and Strusevich, 2009, p. 1
[29] Harjunkoski and Grossmann, 2001, p. 1649
[30] Zhao and Yang, 2016, p. 3418
[31] Li et al., 2023, p. 1-2
[32] Tang and Wang, 2010, p. 104-116 and Pan et al., 2019, p. 327-348
[33] Sarna, 2014
[34] Tasgetiren et al., 2007, p. 271-273

ferences between annealing temperatures of adjacent coils were considered, stringers and tardiness were not. Since DDEs are difficult to apply to real-life problems, the algorithms' usefulness may be in question.[35]

Evolutionary algorithms also tend to focus on local optima, which can be less optimal than a global optimum.[36] The so-called Tabu search solves this problem and was used for optimization in continuous galvanizing lines.[37] Gao et al. (2008) claimed their respective approaches to be very effective as they were able to solve instances with up to 100 coils in only some seconds. But it has to be noted that some aspects like tardiness were not considered. A study by Pan et al. (2017) aimed to minimize the total weighted completion times on a parallel CAL with up to 18 different lines and 120 coils and achieved optimal solutions in under ten minutes.[38]

One similarity of the previously discussed studies is the deterministic characteristic of the used models. Therefore, some authors used dynamic approaches like a multi-agent system (MAS) to optimize dynamic scheduling problems, which consider randomness.[39] A MAS uses artificial intelligence and multiple agents or perspectives to solve a problem to provide flexible solutions.[40] Iannino et al. (2021) used both deterministic and dynamic models for scheduling. The authors used three different approaches to optimize a day's schedule with around 2100 coils in several iterations, with the objective to improve scheduling flexibility. For short-term planning with unstable circumstances, they used a MAS. The second approach was a deterministic, mixed integer linear program (MILP), as is used in this study. The third and last approach utilized a continuous flow model (CFM) for long-term scheduling, which uses a simplified model to schedule the manufacturing process over various time periods.[41] Iannino et al. (2021) summarized that all three approaches have their advantages and disadvantages and that they should be used complementarily rather than substitutionally to utilize each advantage in the right circumstances.

The second most important study about scheduling for this thesis is by Mujawar et al. (2012). The authors proposed a MILP for the minimization of both stringers and tardiness in terms of total time overdue per coil. Due to technical limitations, they were only able to solve the model with up to 15 coils, with a solving time of close to three hours. Because of these limitations, they leveraged two different heuristics, which were able to yield feasible solutions for up to 150 coils, but could only minimize the number of stringers introduced in the schedule and disregarded the tardiness.[42]

Nonetheless, the proposed mathematical model mimicked the real-world process to a higher degree than others. Because of this, Wegel et al. (2024) decided to base their study on the model proposed by Mujawar et al. (2012) and develop it further for better performance in the solving process.

3. Optimization model for Parallel Heterogeneous Annealing Line Scheduling

3.1. Problem description

As mentioned in section 2.1, the continuous annealing process is very complex, and many parameters have to be accounted for. One such parameter, that was briefly mentioned before, is the individual due date of each coil. These are necessary, to schedule the further downstream production stages, like the galvanization.[43] These schedules only have limited flexibility and to not put them at risk, the maximum number of delays has to be bound. This limitation is implemented through a service level, which value is relative to the number of coils processed. Thus, a greater instance is granted with a higher service level than a small or medium one.

Until now, only the specific characteristics of the coils and their processing modes have been considered. But the lines they are processed on have different characteristics themselves. Some lines may only be able to process coils with certain thicknesses and widths, while others may be able to process all coils, regardless of their characteristics. When creating a schedule for a continuous annealing line, the manufacturer thus also has to consider which coil can be processed on which of the parallel heterogeneous lines, therefore decreasing the amount of planning flexibility.

This results in a schedule that has to consider the different characteristics of the coils, their desired specifications in terms of strip speed and annealing temperature, the compatibility of said characteristics and processing modes of consecutive coils, as well as their compatibility with each processing line and their due dates, while it also has to comply with the service level and aims to minimize the number of used stringers. To be able to create such a complex schedule, Wegel et al. (2024) set four assumptions, regarding the release dates, internal service level, costs, and operating conditions.

To reduce complexity, it is first assumed that every coil is available and waiting for processing at the beginning of the schedule. Therefore, each coil could be the first to be processed in the schedule since they have no release dates.

The second assumption concerns the previously described internal service level. As already mentioned, the service level bounds the absolute number of delayed coils to a certain value that is relative to the instance size. But the internal service level also bounds the maximum delay that can occur in the schedule, since all coils have to be processed during it. This maximum delay depends on the maximum amount of time needed to process all coils in one schedule.

[35] Zhao and Yang, 2016, p. 3418
[36] Mirjalili and Gandomi, 2023, p. 393
[37] Gao et al., 2008, p. 1829-1833
[38] Zhang and Yang, 2014, p. 800-802
[39] Cowling et al., 2004, p. 178-188 and Ouelhadj et al., 2004, p. 161-172
[40] Balaji and Srinivasan, 2010, p. 1-2
[41] Iannino et al., 2021, p. 620-630
[42] Mujawar et al., 2012, p. 440-444
[43] Zhao and Yang, 2016, p. 3418

Table 1: Notation of the mathematical model.

Symbol	Meaning
	Indices and index quantities
$i, j \in 1, ..., \mathcal{N}$	Coils
$k \in 1, ..., \mathcal{K}$	Processing lines
$m \in 1, ..., \mathcal{M}_{ik}$	Feasible processing modes of coil i on line k
$n \in 1, ..., \mathcal{M}_{jk}$	Feasible processing modes of coil j on line k
	Parameters
d_i	Due date of coil i
p_{ikm}	Processing time of coil i on line k in mode m
α	Service level
M	Maximum duration of the schedule
c_{ijkmn}	Costs of adding a stringer between coils i in mode m and j in mode n on line k
t_{ijkmn}	Additional processing time of a stringer between coils i in mode m and j in mode n on line k
	Decision variables
$y_{ijkmn} \in \{0, 1\}$	Binary variable with value 1, if coils i in mode m and j in mode n are processed in sequence on line k, else 0
$\delta_{ijk} \in \{0, 1\}$	Binary variable with value 1, if coils i and coil j are processed in sequence on line k, else 0
$x_{ikm} \in \{0, 1\}$	Binary variable with value 1, if coil i is processed in mode m on line k, else 0
$s_i \geq 0$	Start date of coil i
$z_i \in \{0, 1\}$	Binary variable with value 1, if coil i is delayed, else 0

The third assumption made sets the costs of all coils to a fixed value. The reasoning behind this is that only the costs caused by the introduction of stringers should be considered since we cannot avoid the costs caused by the processing of coils. Due to CALs being quite stable and completely automated, Wegel et al. (2024) also assume that no randomness occurs during the process and therefore propose a deterministic model, in which all parameters are known a priori. This is the fourth and last assumption regarding the mathematical model.

Based on these assumptions and the aspects mentioned previously, the authors build a deterministic optimization model that aims to create an optimal, cost minimizing, and tardiness-bound schedule for parallel heterogeneous annealing lines, which is based on the model by Mujawar et al. (2012)[44] In the following sections, the notation of this mathematical model and the model itself will be explained, which will be followed by further extensions to it.

3.2. Notation

This section presents the notation of the mathematical model proposed by Wegel et al. (2024). The coils of flat steel are denoted by $\mathcal{N} = \{1, ..., N\}$ and the parallel continuous annealing lines they are processed on by $\mathcal{K} = \{1, ..., K\}$. Every coil $i \in \mathcal{N}$ has a specific due date $d_i > 0$. The maximum amount of acceptable delays, the service level, is denoted by $\alpha \in \mathbb{N}$.

Each coil i also has a set amount of different processing modes $\mathcal{M}_{ik}$ on each line $k \in \mathcal{K}$. The feasible process mode m of coil i on line k includes a specific annealing temperature and strip speed. If coil i can be processed in a specific mode on line k depends on the characteristics of the coil i and of the line k, as well as on the parameters of the processing mode itself. It is therefore possible, that no mode for the processing of coil i on line k exists. If coil i can be processed on line k in mode m, the processing time p_{ikm} can be derived.

Furthermore, coil i on line k could be succeeded by coil j, with $j \in \mathcal{N}$, in the feasible processing mode n, with $n \in \mathcal{M}_{jk}$. In this case, the compatibility of coil i in mode m and coil j in mode n regarding their width, thickness, annealing temperatures, and strip speeds has to be reviewed. If the coils and modes are incompatible with each other, a stringer has to be added between them. The additional costs caused by introducing a stringer into the schedule between coil i in mode m and coil j in mode n on line k are expressed by c_{ijkmn}, while the additional processing time is contained in t_{ijkmn}. If the coils and their respective modes are compatible, the additional costs and processing time will equal zero, since no stringer has to be introduced.

The processing schedule will be defined by the following decision variables. Each coil's start time is represented by $s_i \geq 0$, while its processing line and mode in the optimized

[44] Mujawar et al., 2012, p. 440

schedule are indicated by the binary variable $x_{ikm} \in \{0,1\}$ taking the value of one if coil i is being processed on line k in mode m or zero, if not. If the completion of coil i is behind the scheduled due date d_i, the binary variable $z_i \in \{0,1\}$ will equal one or zero, if coil i's processing was finished in time. The sequence of the schedule is indicated by the binary variable $\delta_{ijk} \in \{0,1\}$, with the value of one if coil i precedes coil j on line k or zero if this is not the case.

Additional information regarding their respective modes m and n is provided by the binary variable $y_{ijkmn} \in \{0,1\}$ with the value of one if coil i is being processed in mode m and is succeeded by coil j in mode n on line k or zero, if not.

3.3. Mathematical model

$$\min_{z,s,x,\delta,y} Z = \sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{N}}\sum_{k\in\mathcal{K}}\sum_{m\in\mathcal{M}_{ik}}\sum_{n\in\mathcal{M}_{jk}} c_{ijkmn}\cdot y_{ijkmn}$$

s.t.

$$s_i + \sum_{k\in\mathcal{K}}\sum_{m\in\mathcal{M}_{ik}} p_{ikm}\cdot x_{ikm} \le d_i + z_i\cdot M \qquad \forall i\in\mathcal{N}, \quad (1)$$

$$s_i + \sum_{k\in\mathcal{K}}\sum_{m\in\mathcal{M}_{ik}}\sum_{n\in\mathcal{M}_{jk}} (p_{ikm}+t_{ijkmn})\cdot y_{ijkmn} \le s_j + M\cdot(1-\sum_{k\in\mathcal{K}}\delta_{ijk}) \qquad \forall i,j\in\mathcal{N}, \quad (2)$$

$$\delta_{ijk} = \sum_{m\in\mathcal{M}_{ik}}\sum_{n\in\mathcal{M}_{jk}} y_{ijkmn} \qquad \forall i,j\in\mathcal{N}, \forall k\in\mathcal{K}, \quad (3)$$

$$\sum_{j\in\mathcal{N}}\sum_{n\in\mathcal{M}_{jk}} y_{ijkmn} \le x_{ikm} \qquad \forall i\in\mathcal{N}, \forall k\in\mathcal{K}, \forall m\in\mathcal{M}_{ik}, \quad (4)$$

$$\sum_{i\in\mathcal{N}}\sum_{m\in\mathcal{M}_{ik}} y_{ijkmn} \le x_{jkn} \qquad \forall j\in\mathcal{N}, \forall k\in\mathcal{K}, \forall n\in\mathcal{M}_{jk}, \quad (5)$$

$$\sum_{k\in\mathcal{K}}\sum_{j=1}^{\mathcal{N}+1}\delta_{ijk} = 1 \qquad \forall i\in\mathcal{N}, \quad (6)$$

$$\sum_{k\in\mathcal{K}}\sum_{i=0}^{\mathcal{N}}\delta_{ijk} = 1 \qquad \forall j\in\mathcal{N}, \quad (7)$$

$$\sum_{j=1}^{\mathcal{N}+1}\delta_{0jk} = 1 \qquad \forall k\in\mathcal{K}, \quad (8)$$

$$\sum_{i=0}^{\mathcal{N}}\delta_{i(\mathcal{N}+1)k} = 1 \qquad \forall k\in\mathcal{K}, \quad (9)$$

$$\sum_{j=0}^{\mathcal{N}}\delta_{jik} = \sum_{j=1}^{\mathcal{N}+1}\delta_{ijk} \qquad \forall i\in\mathcal{N}, \forall k\in\mathcal{K}, \quad (10)$$

$$\sum_{i\in\mathcal{N}} z_i \le \alpha, \quad (11)$$

$$s_i \ge 0 \qquad \forall i\in\mathcal{N}, \quad (12)$$

$$x_{ikm}, \delta_{ijk}, y_{ijkmn}, z_i \in \{0,1\} \qquad \forall i,j\in\mathcal{N}, \forall k\in\mathcal{K}, \forall m\in\mathcal{M}_{ik}, \forall n\in\mathcal{M}_{jk}. \quad (13)$$

The main objective of this mathematical model is the minimization of costs caused by the introduction of stringers. Therefore, the objective function minimizes the sum of products of c_{ijkmn} and y_{ijkmn}. By multiplying c_{ijkmn} and y_{ijkmn} the cost of a stringer is only regarded if coil i in mode m and coil j in mode n on line k are not compatible with each other and are processed in sequence in the optimized schedule, thus requiring a stringer. If they are not processed in sequence or if they are compatible with each other or both, the value of the product will equal zero since y_{ijkmn} or c_{ijkmn} or both will equal zero, respectively. By taking the sum over all coils, lines and modes, all introduced stringers in the schedule are considered.

To mimic the real-world process and maintain consistency, several constraints are necessary and will be discussed in the following. Constraints (1) and (2) assure time consistency, while constraints (3)-(5) ensure the consistency of decision variables that indicate the same information. The consistency of the processing sequence in the schedule is established by the constraints (6)-(10). These are followed by three additional constraints, that regulate general aspects of the mathematical model and the schedule.

The first constraint (1) takes the tardiness of the schedule into account. The time of completion of coil i is calculated by adding its processing time p_{ikm} in mode m on line k to its start date s_i. This completion date has to be smaller or equal to its due date d_i for the coil to be finished processing in time. In this case, z_i would equal zero due to the restriction of the number of delayed coils by the service level and the possibility that the delay could be used somewhere else in the schedule to further minimize the number of introduced stringers. But if the completion date is after the due date, z_i has to equal one to not violate the equation, since the time of completion of coil i will be greater than its due date d_i. Additionally, the product of z_i and M assures that coil i has to be processed sometime in the schedule since M is defined as $max_k(\sum_{i\in\mathcal{N},\mathcal{M}_{ik}\neq\emptyset} max_{j,m,n}(p_{ikm}+t_{ijkmn}))$, which is the maximum amount of time necessary to process all coils in the schedule. This constraint has to be set for every coil, as every coil has a due date and could be delayed.

Constraint (2) considers the time sequences of processing in the schedule. The left side represents the earliest time a succeeding coil j could be processed after the processing of the preceding coil i was finished. This is achieved by adding the processing time p_{ikm} of coil i in mode m on line k and the possible additional processing time of a stringer t_{ijkmn}, in case of incompatibility of coil i in mode m and coil j in mode n, to the start date of coil i. Processing of the succeeding coil j can only start after the processing of coil i or that of

the stringer is completed. Thus, s_j has to be greater or equal to that time of completion. In the case that coils i and j are processed in sequence on line k, δ_{ijk} would equal one, and thus, M would not be added to the right-hand side. But as this constraint holds for all coils, it could compare two coils with each other, that are not processed in sequence. To prevent this comparison to have an effect on the solution, M is added to s_j, since in this case, δ_{ijk} will equal zero. Thus, the right side will always be greater than or equal to the left side since $M \leq s_j + M$ and $s_i + (p_{ikm} + t_{ijkmn}) \cdot y_{ijkmn} \leq M$ in every scenario, because of the nature of M described previously.

Constraint (3) assures consistency throughout the variables δ_{ijk} and y_{ijkmn}. Both variables indicate the sequence of coils i and j, as well as the line k they are processed on, and therefore should be equal to each other with the same coils i, j and line k. Since y_{ijkmn} also indicates the processing modes of both coils and cannot represent a binary value without this information, the sum of y_{ijkmn} over all modes has to be taken for this constraint to be effective. This constraint also reassures, that the coils are only processed in one mode and not in multiple, hence in this case, two y_{ijkmn} would equal one with the same coils and line. Therefore, δ_{ijk} would have to equal two, which is impossible due to the binary nature of δ_{ijk}.

Constraint (4) contributes to the decision variable consistency as well. Both variables y_{ijkmn} and x_{ikm} indicate the processing line k and the processing mode m of coil i. Hence, they should be equal to each other for every coil, mode, and line. But in the case that coil i is the last coil that is being processed on line k, y_{ijkmn} would equal zero since y_{ijkmn} does not take the coils $i \in \{0, \mathcal{N}+1\}$ into account and could therefore not be succeeded by another coil j. To not violate this restriction in this case, y_{ijkmn} must only be less or equal to x_{ikm}. Therefore, with x_{ikm} equalling one, y_{ijkmn} could be either one or zero. Furthermore, by taking the sum of y_{ijkmn} over all coils and modes, the constraint prevents coil i from being succeeded by multiple different coils in multiple different modes and limiting this number to the value of x_{ikm}, which is either one or zero due to its binary nature.

Therefore, coil i can only be succeeded by a single coil j in a single mode n on line k, if i is being processed on line k in mode m or by none if it is not. This holds for all coils, lines, and modes.

The following constraint (5) is essentially the same as (4) but for the succeeding coil j. Hence, it prevents coil j from succeeding multiple coils in multiple modes by limiting this number to x_{jkn}. It has to be noted, that the indices of x_{jkn} in this constraint differ from the indices of x_{ikm} in the other constraints since this constraint should only restrain the succeeding coil j. Additionally, y_{ijkmn} must only be less or equal to x_{jkn}, since all y_{ijkmn}'s would be zero if j would be the first coil to be processed on line k, which would violate the constraint. This is because this constraint only accounts for coils in $\mathcal{N}$ and thus the first coil on line k cannot be preceded. This restriction applies to all succeeding coils on every line in every processing mode.

Constraint (6) restricts the number of coils that succeed coil i on any line to one. This prevents coil i from being processed and preceded by several coils on different lines. If this would be the case, the sum of δ_{ijk} over all lines and coils in $[1, \mathcal{N}+1]$ would be greater for coil i than one, thus violating the restraint. If coil i is the last coil to be processed on line k, it could not be succeeded by any coil, resulting in δ_{ijk} equalling zero. Therefore, a virtual coil $\mathcal{N}+1$ has to be introduced so that in the case mentioned above, coil i would not violate this constraint. This restriction applies to every coil.

Constraint (7) is similar to constraint (6), as this constraint restricts the number of preceding coils of coil j on all lines to one and therefore prevents coil j from being processed and preceded more than once during the schedule. The extension of $\mathcal{N}$ by zero can be explained with the constraint (8). Here, i is set to the value zero. This virtual coil zero marks the beginning of the schedule on every line k. Since this coil is not being processed, it is not included in most other constraints. But since coil zero is the first on every line k, it has to be succeeded by one coil $j \in [1, \mathcal{N}+1]$. Therefore, constraint (8) ensures that one coil j is the first one to be processed on line k. No real coils are processed on line k during the schedule if coil $\mathcal{N}+1$ succeeds coil zero. This holds for every line because every line has to have a virtual first coil that is being succeeded by another coil, virtual or real.

Constraint (9) on the other hand ensures that one coil $i \in \{0, \mathcal{N}\}$ has to be the last real coil to be processed on line k. To mark this coil as the last coil to be processed on line k, it will be succeeded by another virtual coil $\mathcal{N}+1$. As on every line k there has to be a last coil to be processed, this holds for every line.

The last sequence consistency constraint, constraint (10), assures that every real coil i has to have a predecessor and successor coil j on the line k it is being processed on. But it also accounts for the extreme case, that a processing line k processes no real coil, which is why it considers coil zero as a predecessor on the left-hand side and coil $\mathcal{N}+1$ as a successor on the right-hand side. Additionally, this considers that one real coil has to be the first and another has to be the last coil if real coils are processed on the line, as described in constraints (8) and (9). This restriction applies to every line.

Hence, the introduction of the virtual coils zero and $\mathcal{N}+1$ prevents δ_{ijk} from equalling zero in the cases that coil j is the first or coil i is the last coil to be processed on line k, which would violate the sequence consistency constraints.

The eleventh constraint (11) restricts the maximum amount of delayed coils to the service level α. As this applies to the entire schedule, the sum of z_i over all coils has to be taken. Since processing only starts at the beginning of the schedule, no start date s_i can take a value of less than zero. This is ensured by constraint (12). As every coil needs a start date, since every coil has to be processed at some point during the schedule, this holds for every coil. The last constraint (13) ensures that the binary variables x_{ikm}, δ_{ijk}, y_{ijkmn}, and z_i only equal the binary values of zero or one and also indicates which index is the element of what parameter.

3.4. Expansions of the model

3.4.1. Release dates

The first expansion tackles the assumption of no coil-specific release dates made by Wegel et al. (2024), which was explained in section 3.1. To mimic a more realistic process, coil-specific release dates were added to the model with the new parameter r_i. This parameter holds the specific release date of each coil i, with the release dates being distributed with the same distribution as the urgency of the coils, but with zero as the lower and a as the upper bound. Therefore, in the case of $Very\ high$ UC, the release dates are distributed via the uniform distribution $U(0, a)$. The UC and other factors will be explained in section 4.1 For the release dates to be considered in the model, the following constraints had to be added:

$$s_i \geq r_i \qquad \forall i \in \mathcal{N} \quad (14)$$

$$r_i \geq 0 \qquad \forall i \in \mathcal{N} \quad (15)$$

Constraint (14) restricts the start time of coil i, s_i, from being earlier than its release date r_i, while (15) prevents the release date from being earlier than the start of the schedule. These hold for all coils.

This is only a minor extension of the model. The two major extensions will be described in the following sections.

3.4.2. Minimization of stringer use and tardiness

Since both tardiness and costs caused by stringer introduction play a pivotal role in the scheduling of parallel heterogeneous annealing lines (PHALs), Wegel et al. (2024) proposed an expansion of the mathematical model. This expansion modifies the objective function in such a way, that the sum of introduced stringers and the absolute number of delays are minimized, as depicted below. Because of this, the service level α and constraint (11) from the mathematical model in section 3.3 are not necessary anymore and are removed. To account for different weights of tardiness and stringer costs, the two sub-objectives are multiplied with new parameters. While r_str sets the weight of the stringer costs, r_tar does the same for the tardiness, expressed by the sum of delayed coils, with $r_str + r_tar = 1$. To achieve this, an additional restriction (16) was introduced. This constraint ensures that the total weight of tardiness and stringer costs does not differ from 100%.

$$\min_{z,s,x,\delta,y} Z = \sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{N}}\sum_{k\in\mathcal{K}}\sum_{m\in\mathcal{M}_{ik}}\sum_{n\in\mathcal{M}_{jk}} c_{ijkmn}\cdot y_{ijkmn}\cdot r_str + \sum_{i\in\mathcal{N}} z_i \cdot r_tar$$

$$r_str + r_tar = 1 \quad (16)$$

3.4.3. Minimization of stringer use and due date deviation

The second major extension of the model is an enhancement of the first extension by not only taking tardiness into account but also the earliness of the schedule. If products are produced before the due date, they have to be stored in warehouses, which leads to so-called inventory holding costs. They are not only caused by the storage and handling of the inventory, but also by insurance and other factors.[45]

Thus, there is also an incentive to minimize earliness in production processes. To minimize both the use of stringers and the deviation from the due date, the objective function had to be altered accordingly.

$$\min_{z,s,x,\delta,y} Z = \sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{N}}\sum_{k\in\mathcal{K}}\sum_{m\in\mathcal{M}_{ik}}\sum_{n\in\mathcal{M}_{jk}} c_{ijkmn}\cdot y_{ijkmn}\cdot r_str + \sum_{i\in\mathcal{N}}(v_i\cdot r_v + e_i\cdot r_e)\cdot r_dd$$

The first part of the objective function is the same as in the first major extension. But the second part introduces two new variables and three new parameters. The two new continuous variables e_i and v_i can be explained by describing the new first constraint:

$$s_i + \sum_{k\in\mathcal{K}}\sum_{m\in\mathcal{M}_{ik}} p_{ikm}\cdot x_{ikm} + e_i = d_i + v_i \quad \forall i \in \mathcal{N} \quad (17)$$

These new continuous variables, with a lower bound of zero, work in such a way, that only one of them can have a value higher than zero for coil i. If coil i in mode m on line k was finished processing before its due date d_i, $s_i + p_{ikm}\cdot x_{ikm} < d_i$. This would violate the constraint since both sides have to be equal to each other. Therefore, e_i will take the value of $d_i - (s_i + p_{ikm}\cdot x_{ikm})$, which represents the positive time left until the due date and therefore the earliness of coil i. In this case, v_i will equal zero, since both e_i and v_i are to be minimized. A v_i greater than zero would lead to $d_i + v_i - (s_i + p_{ikm}\cdot x_{ikm}) > d_i - (s_i + p_{ikm}\cdot x_{ikm})$, therefore increasing both v_i and e_i and thus the objective value.

The opposite case, in which coil i is delayed, is similar. Now, $s_i + p_{ikm}\cdot x_{ikm} > d_i$. To not increase this difference and hence the objective value, e_i will equal zero. Because of the necessary equality of both sides, v_i will equal $s_i + p_{ikm}\cdot x_{ikm} - d_i$, representing the positive time that coil i is delayed.

In the third and last case, in which the processing of coil i is finished processing exactly on time, $s_i + p_{ikm}\cdot x_{ikm} = d_i$. Since e_i and v_i are to be minimized, both variables will equal their lower bound zero. In the altered objective function, e_i and v_i are multiplied with the parameters r_v and r_e, to provide a possibility for weighing the costs of tardiness and earliness. The third and last parameter r_dd weighs the costs of due date deviation against the weight of the stringer costs

[45] Durlinger, 2015

r_str. Since weights cannot differ from 100 %, the following constraints have to hold:

$$r_str + r_dd = 1 \tag{18}$$

$$r_v + r_e = 1 \tag{19}$$

In addition, the variable z_i is removed from the model because absolute tardiness is not relevant anymore.

4. Computational study

4.1. Structure of the study and instances

After describing the industrial process and the mathematical model for creating and optimizing a schedule for CALs in the before sections, this section will outline the structure of the numerical study, which will be followed by the results and the interpretations of it.

The computational study makes use of four main factors proposed by Wegel et al. (2024), with each having five different levels ranging from $Very\ low$ to $Very\ high$, which are depicted in table 2. The heterogeneity of coils (HC) describes the possible range of each characteristic of the coils and their distribution among them. This includes the width, thickness, and length, as well as the mid-temperature and mid-speed, and the symmetric intervals around them. These will be used to assign physical properties to each coil during the computational study, as well as to define the feasible processing modes. The distributions of coil characteristics differ in the five levels. $TR(a, b, b)$ and $U(a, b)$ represent a triangular and uniform distribution, respectively.

As an example, the width in the $Very\ high$ level of HC will be distributed by the uniform distribution $U(60, 100)$, since 60 cm is the lower bound and therefore equal to a and b, the upper bound, equals 100 cm. Hence, the width of each coil will be between 60 cm and 100 cm, with all values having an equal probability. This aspect of the uniform distribution increases the heterogeneity and difficulty of finding a stringer-free solution since the characteristics will be evenly spread between a and b, thus decreasing the probability of compatibility between different coils. With a triangular distribution $TR(a, b, b)$, however, most values will be close to b, hence increasing the probability of compatibility. This is the reason the distribution changes from a triangular to a uniform distribution with increasing intensity of HC.

Each coil receives a mid-temperature and -speed through the distribution of $midTemp$ and $midSpeed$ in steps of ten °C and ten m/min respectively, which represent the centers of symmetric intervals. The widths of these intervals are determined by $\Delta temp$ and $\Delta speed$. The ranges of these intervals indicate the different feasible annealing temperatures and strip speeds, under which the coil can be processed. A combination of feasible annealing temperature and strip speed is a feasible processing mode for that specific coil. In order to limit the number of processing modes, to 48 in total, the annealing temperature and strip speed in processing modes only vary in increments of ten °C and 20 m/min, respectively. To give an example, coil i in the *Basecase* scenario has a mid-temperature of 680 °C and a mid-speed of 510 m/min. Thus, the intervals with $\Delta temp = 20$ and $\Delta speed = 20$ are [670 °C; 690 °C] and [500 m/min; 520 m/min]. The resulting feasible processing modes for coil i are [(670, 500), (670, 520), (680, 500), (680, 520), (690, 500), (690, 520)], with the first entry being the annealing temperature and the second being the strip speed. The maximum number of feasible processing modes per coil is therefore six.

This holds for every level of HC, except for the $Very\ high$ level, due to the narrower interval widths. In this case, the annealing temperature and strip speed can only vary by + ten °C and + ten m/min, respectively. The intervals in the example from above are therefore [680 °C; 690 °C] and [510 m/min; 520 m/min], resulting in only two possible processing modes [(680, 520), (690, 520)]. Hence, each coil has two to six different processing modes, depending on the level of HC. An extension to the HC is the distribution of the length of coils. This results in far more heterogeneous processing times p_{ikm} since the processing time of coil i on line k in processing mode m is defined as the length of coil i divided by the strip speed of mode m. This minor extension will be included in all models used in the computational study.

The urgency of the coils (UC) defines the size of the time horizon that is used for the distribution of the individual due dates. As with HC, lower levels distribute the due dates via a triangular distribution, while higher levels distribute them with a uniform distribution or a mixture of both, thus spreading the due dates more evenly on the time horizon. Adding to this, the uniform distribution distributes more due dates close to the start of the schedule than the triangular distribution. By additionally decreasing the value of a by increasing the UC, the earliest due date will also be decreased, since both the triangular distribution $TR(a, b, b)$ and the uniform distribution $U(a, b)$ are using a as the earliest possible due date. These factors restrict the feasible solution space and could thereby be detrimental to finding an optimal, stringer-free schedule. The latest possible due date b is a function of $\mathcal{N}$, which is defined as $f(\mathcal{N}) = avg_proc_time \cdot \mathcal{N}/\mathcal{K}$, with avg_proc_time being the sum of coil i's length divided by the strip speeds of feasible process modes $\mathcal{M}_{ik}$ over all coils and processing lines, which is then divided by the product of $\mathcal{N}$ and $\mathcal{K}$. That the values of a and b both depend on the number of coils $\mathcal{N}$ assures that the time horizon is adequate to the instance size and does not lead to infeasibility with a higher number of $\mathcal{N}$.

Process flexibility (PF) defines the compatibility between consecutive coils in terms of their thickness, width, annealing temperature, and strip speed. Higher PF allows for a higher difference in these characteristics, thus increasing the scheduling flexibility. To give an example, the difference in widths between two consecutive coils in the *Basecase* scenario can be up to five cm, without having to introduce a

[46] Wegel et al., 2024, p. 20

Table 2: Parameterization of the different factors and levels.[46]

Factor \ Level	Very low	Low	Basecase	High	Very high
Heterogeneity coils					
distribution	TR(a,b, (a+b)/2)	TR(a,b, (a+b)/2)	50%TR(a,b, (a+b)/2), 50%U(a,b)	U(a,b)	U(a,b)
$width(a,b)$ [cm]	70, 90	60, 100	60, 100	60, 100	60, 100
$thickness(a,b)$ [cm]	0.15, 0.35	0.1, 0.499	0.1, 0.499	0.1, 0.499	0.1, 0.499
$length(a,b)$ [km]	15, 16	14, 17	14, 17	14, 17	14, 17
$midTemp(a,b)$ [°C]	690, 720	680, 730	680, 730	680, 730	680, 730
$\Delta temp$ [°C]	20	20	20	20	10
$midSpeed(a,b)$ [m/min]	520, 580	510, 590	510, 590	510, 590	510, 590
$\Delta speed$ [m/min]	20	20	20	20	10
Urgency of coils					
distribution	TR(a,b,b)	TR(a,b,b)	TR(a,b,b)	50%TR(a,b, (a+b)/2), 50%U(a,b)	U(a,b)
a [min]	$\lvert\mathcal{N}\rvert \cdot 2$	$\lvert\mathcal{N}\rvert \cdot 1.5$	$\lvert\mathcal{N}\rvert$	$\lvert\mathcal{N}\rvert$	$\lvert\mathcal{N}\rvert$
b [min]	$f(\mathcal{N})+a$	$f(\mathcal{N})+a$	$f(\mathcal{N})+a$	$f(\mathcal{N})+a$	$f(\mathcal{N})+a$
Process flexibility					
$\Delta width$ [cm]	>5	>5	>5	>10	>15
$\Delta thickness$ [mm]	>0.05	>0.1	>0.1	>0.1	>0.2
$\Delta temp$ [°C]	>0	>0	>10	>10	>20
$\Delta speed$ [m/min]	>0	>20	>20	>20	>20
Stringer processing					
$time$ [min]	0	1.5	3	10	15

stringer between them. If the difference extended five cm, a stringer would have to be added to bridge the differences. Thus, higher PF minimizes the need for stringers and vice versa. Please note, that the length has no PF since it does not have an impact on the welding or process mode compatibility.

The stringer processing time (SPT) defines the amount of time needed to process a stringer if introduced into the schedule. A high SPT diminishes flexibility in terms of time constraints and tardiness. There were no different stringer costs added to the base model since these are the only costs that are being considered. Therefore, different stringer costs would not affect problem-solving in any way. However, the extensions include relative stringer costs through r_str, since the minimization of stringers is only one of two different sub-objectives.

Since the optimization model was built for a PHAL, three lines with different specifications were chosen for the numerical studies, if not stated otherwise. For this study, the heterogeneity of the lines will be limited to the width of the coils the line can process. The first line can process all coils, regardless of their width. Only narrow coils can be processed on the second line, while only wide ones can be processed on the third one. This follows the line specifications selected by Wegel et al. (2024), which were adopted for this study, as well as the maximum solving time of 720 seconds. This time limit is not a hard limit, hence some optimizations might slightly exceed it. The exact line specifications depend on the chosen level of HC. They are configured such they can only process coils that are five cm above or below the width that represents the center of the width range. As an example, with $High$ HC, the second line can only process coils with a width of less than or equal to 75 cm.

First, the tuning parameterization will be discussed. Tuning influences the behavior of the solver and thus could increase its performance.

The second part consists of a study on different instance sizes and the *Basecase* scenario for each factor. For each instance size, the service level will vary between low, medium and high, with low representing a service level of zero %, medium representing ten % and high 20 %. Each optimiza-

tion in the numerical study will be conducted with the same ten seeds, to ensure a certain degree of comparability between the different instant sizes and sections.

A seed determines the outcome of randomness, like the distribution of coil characteristics or due dates. Therefore, by increasing the number of coils, new coils will be added to the already existing ones. Hence, two calculations with the same parameters and the same seed yield nearly the same result, but the optimization itself could still be affected by minor randomness.[47] This assures that differences in solution characteristics between two calculations with different parameters, i.e., different service levels, are mostly caused by the different parameters and not by randomness.

This section of the study aims to uncover possible trends that occur with certain parameterizations. The solutions and performance achieved with the default solver will be compared with the tuned solver. At the end of this part, an instance size will be chosen that will be used for the rest of the study.

For the third part of the numerical study, the Best-case and Worst-case scenarios are defined by using the different factors in table 3.

Table 3: Parameterization of the Worst- and Best-case scenario.

Scenario	HC	UC	PF	SPT
Best-case	Very low	Very low	Very high	Very low
Worst-case	Very high	Very high	Very low	Very high

The results of these extreme cases will be compared with the *Basecase* scenario to determine the impact of the four factors. Additionally, the factors PF and HC will be altered in the Best- and Worst-case scenarios, respectively, to further investigate their influence. The levels used for the factors are $Very\ low$, *Basecase* and $Very\ high$.

The fourth section will explore the impact of changing factor and model specifications on the solution. First, the number of processing lines in the *Basecase* scenario will be altered, as depicted in table 4.

Table 4: Parameterization of processing line altering cases.

Case	Alteration
Addition	Addition of processing line, that can process all coils
Removal I	Removal of processing line, that can only process wider coils
Removal II	Removal of processing line, that can only process narrower coils

[47] Miltenberger, 2023b

The results will be compared to the *Basecase* scenario with the default number of processing lines, with emphasis on the number of introduced stringers and tardiness. Secondly, the a and b parameters used in the UC will be reduced to study the impact of earlier due dates on the feasibility of the scheduling. For each of the cases, the *Basecase* scenario with a $Very\ high$ UC will be used.

The last section compares some aspects of the solutions of the original model with the solutions of the expanded models. The first expanded model to be studied will be the model which aims to minimize both the number of stringers and absolute tardiness. In addition to varying weights of stringer and tardiness costs, the UC will be altered to study its influence on the solution. This results in the cases in table 6 that will be explored with a $Very\ low$, *Basecase* and $Very\ high$ UC.

These cases are chosen to accommodate for different weighing preferences and since it is assumed, that under no circumstances one of the costs could be neglected. Thus, no case only considers stringer or tardiness costs.

The second expanded model, which aims to minimize stringer, earliness and tardiness costs, will also be studied using the different cases in table 7. Different stringer and due date deviation weights will be combined with different earliness and tardiness weights, which are the components of the due date deviation.

The entire computational study consists of 1180 individual runs, with an approximate total computation time of 140 hours and was conducted on a computer with a Ryzen 7 3700X CPU, 16 GB of 3200MHz memory, and Windows 10. The CPU frequency could have differed due to changing room temperature and longer computation sessions, thus slightly affecting the performance of some optimizations.

4.2. Results and interpretations

4.2.1. Tuning of the solver

In the following, the tuning for each instance size will be briefly explained, which was defined through trial and error. These runs and optimization times were not included in the above-mentioned numbers of 1180 runs and 140h. It has to be noted, that the tuning parameters of a lower instance size are also applied to higher instance sizes, if not stated otherwise.

The two smallest instance sizes ten and 20 coils received only minor tuning. The so-called $PreSolve$ was set to two, which is an aggressive setting. $PreSolve$ tightens the model before solving it by, as an example, removing redundant and inactive constraints, thus increasing the probability of finding a feasible solution and the performance of the solver in general. The duration of $PreSolve$ was cut short with the $PrePasses$ parameter since smaller instances already have smaller models and thus find feasible solutions fast, as will be explained later in the second section. Another parameter, $Cuts$, was set to conservative, which behaves like $PreSolve$ by tightening the model. The difference between $PreSolve$

Table 5: Parameterization of due date altering cases.

Case	Alteration	a [min]	b [min]
100 %	Very high UC	40	1306
75 %	Very high UC and reduction of a and b by 25%	30	979.5
50 %	Very high UC and reduction of a and b by 50%	20	653
25 %	Very high UC and reduction of a and b by 75%	10	326.5

Table 6: Parameterization of cases with different stringer and tardiness weights.

Case	Stringer weight [%]	Tardiness weight [%]
1	80	20
2	65	35
3	50	50
4	35	65
5	20	80

Table 7: Parameterization of cases with different stringer, due date deviation, earliness and tardiness weights.

Case	Stringer:Due date deviation weight [%]	Earliness weight [%]	Tardiness weight [%]
	65:35		
1		80	20
2		65	35
3		50	50
4		35	65
5		20	80
	50:50		
6		80	20
7		65	35
8		50	50
9		35	65
10		20	80
	35:65		
11		80	20
12		65	35
13		50	50
14		35	65
15		20	80

and Cut is that $Cuts$ tightens the model during the solving process and not before it, as $PreSolve$ does.[48]

The $PrePass$ parameter will be set to automatic or negative one when solving instances with more than 20 coils, allowing for a longer $PreSolve$ and thus more tightening. Additionally, Gurobi-intern heuristics are maximized through the $heuristics$ parameter. Gurobi has many types of complex heuristics that can improve the ability to find a feasible solution significantly. By increasing the $heuristics$ parameter, the solver will spend more time on heuristics. The reason for this aggressive parameterization will be discussed in the next section regarding figure 2. Another tuning parameter, the $MIPFocus$, was modified with instances of 50 coils and higher. It influences, how much time will be spent on finding feasible solutions and on the optimality of the solutions. The $MIPFocus$ was parameterized to one to focus on finding feasible solutions rather than proving their optimality with the MIPGap. The MIPGap is a qualifier of a feasible solution, that only exists in Gurobi with MIPs. It is a relative value that represents the minimal optimality of the solution and is defined as $|ObjBound - ObjVal|/|ObjVal|$. To explain the MIPGap, the solving process of Gurobi has to be briefly explained first.[49]

For MILPs, Gurobi uses the so-called branch and bound approach. The solving process starts by solving the model without its integrality restrictions. If the resulting solution does not violate any restrictions, the solving stops since the solution is optimal. But in most cases, some variables of the solution violate constraints. In this case, one of these variables will be chosen as a branching variable and the infeasible solution becomes a node, from which branches extend. These branches are MIPs, that are more restricted in the possible values of the branching variable. After solving these MIPs, they can either generate a feasible solution or become a new node. This process will then be repeated until either a feasible solution was found or it was proven that there are no feasible solutions in this branch. This definitive end of a branch is called a leaf, but all yet unexplored branches are also leaves. If the first or a new best solution was found, it becomes the incumbent solution. The objective value of this incumbent solution is the $ObjVal$ mentioned above in the formula for the MIPGap. The $ObjBound$, also called the best bound, is the lowest or highest value that the solver approximates can be achieved by taking the lowest or highest

[48] Gurobi, n.d.-c and Gurobi, n.d.-b
[49] Miltenberger, 2023a and Gurobi, n.d.-b

Table 8: Tuning parameterization for each instance size.

# coils / Parameter	10	20	30	40	50	60	70
Cuts	1	1	1	1	1	2	2
PrePasses	1	1	-1	-1	-1	-1	-1
PreSolve	2	2	2	2	2	2	1
Heuristics	-	-	1	1	1	1	1
MIPFocus	-	-	-	-	1	1	1
NodefileStart	-	-	-	-	-	-	0.5
Threads	-	-	-	-	-	-	8

value of all the different leaves, depending on the direction of the optimization. For a minimization model, which is used in this study, it will take the lowest value. This best bound can change throughout the solving process through new information about the feasibility of a solution with the best bound as an objective value. As an example of a minimization problem, if the solver notices that the best bound cannot be achieved in a feasible solution, it could raise the best bound to the second-lowest value.[50]

The MIPGap, therefore, is the absolute difference between the objective value of the best already found solution and the objective value of the best solution the solver approximates could be possible, relative to the former. Hence, the MIPGap represents the maximum relative improvement the solver approximates could be possible over the current incumbent solution.[51] Another function of MIPGap is, that the solver will terminate the solving process if the MIPGap of one solution is equal to or lower than the MIPGap that was set as a threshold, with the default threshold being close to zero %, or if the difference between the lower and upper bound is less than the MIPGap multiplied with the upper bound.[52]

For optimizations with 60 coils, the $Cuts$ parameter was set to two, aggressive, to improve the tightening of the model. The number of out-of-memory errors increased dramatically with optimizations of instance sizes with 70 coils, which will be discussed later. To avoid this error, the number of CPU threads was limited to eight by using the parameter $Threads$ and the RAM usage was reduced by increasing $NodefileStart$ to 0.5. Higher thread counts generally lead to higher performance, but also demand more memory, which is why the amount of available threads for Gurobi was limited. Additionally, the $NodefileStart$ parameters relieves the RAM by storing data on the disk. Both of these parameters thus decrease RAM usage and out-of-memory errors, while also decreasing performance, which is why they have to be fine-tuned to not be detrimental to the optimization process.[53]

4.2.2. Impact of increasing instance size

Before inspecting all the data from the first section of the numerical studies, figure 2 has to be considered first. The figure depicts the number of optimizations per instance size and per service limit that did not result in a feasible solution.

It can be observed, that this number rises with the increasing instance size and that this trend is weaker in the tuned model and with higher service levels. As previously described, α determines the maximum amount of delayed coils relative to $\mathcal{N}$. Therefore, an optimization with a higher α is less restrained and has higher scheduling flexibility than one with a lower service level. This data suggests, that due to the higher scheduling flexibility, optimizations with higher service levels can solve for a feasible solution more easily. But it is important to note that this data includes three different reasons for the failed optimization.

The first reason is, that the Gurobi solver could not find a feasible solution during the 720 seconds time limit, thus aborting the solving process. As previously described, this is affected by the service level. Most unsuccessful optimizations occurred because of this reason. The second reason was an out-of-memory error. It occurs if the Gurobi solver demands more RAM than is physically available. This terminates the solving process, without finding a feasible solution prior in all cases. This happened only with greater instance sizes, especially with 70 coils. But the two failed optimizations at an instance size of only 20 coils happened due to a different reason. These optimizations failed due to the infeasibility of the model, in both the default and the tuned model. Since this only happened with a low service level, it suggests that these two seeds cannot be optimized for an instance size of 20 coils without delay. The most probable cause is a combination of low compatibility due to high coil characteristic heterogeneity, early due dates d_i, and high coil processing times p_{ikm}.

Regardless of the reason, this results in a decreasing number of data points for greater instance sizes. Hence, the validity of trends observed in this data could be affected by a

[50] de Harder, n.d. and Gurobi, n.d.-c
[51] de Harder, n.d.
[52] Miltenberger, 2023a and Gurobi, n.d.-b
[53] Gurobi, n.d.-b and Gurobi, n.d.-a, p. 821, 807

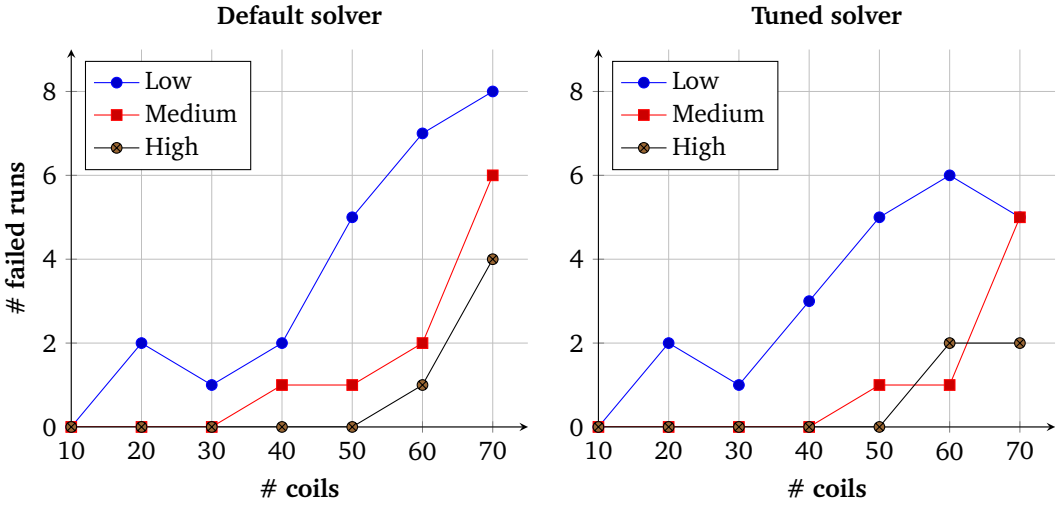


Figure 2: Number of optimizations per $\mathcal{N}$ and service level that yielded no feasible solution with the default and tuned solver.

selectivity effect, in which only optimizations of such seeds yield feasible solutions, that are easier to solve. This could be due to a more homogeneous distribution of coil characteristics, later due dates, and many other factors. Additionally, these seeds do not have to be easier to solve than others for every instance size since additional coils can, as an example, have earlier due dates and thus decrease the scheduling flexibility of the optimization. This notion has to be considered throughout this section of this study.

As depicted in figure 2, optimizations with the tuned solver yielded a feasible solution more often than with the default solver, with 40 failed optimizations with the default and only 33 unsuccessful optimizations with the tuned solver, with a total amount of 210 optimizations per solver. The impact of the tuning seems to rise with the increasing number of coils. This is due to the fact, that with higher instance sizes, more tuning was applied since these had a higher ratio of failed optimizations. After discussing the problem of data density, the characteristics of the solutions will be investigated in the following.

Figure 3 displays that the average number of introduced stringers varies between different instance sizes and service levels but generally seems to decline with the increasing number of coils. The tuning of the solver seems to have no significant effect on the quality of the solution since the objective values yielded by both solvers are mostly similar and only vary by a small amount. This was to be expected because the tuning mainly focussed on improving the solver's performance and not the solutions yielded by it.

But it can be observed, that allowing delays and thus increasing the scheduling flexibility typically results in more optimal solutions regarding stringer introduction. Optimizations with a service level of 20 % and hence the highest scheduling flexibility generally yielded the best solutions concerning stringers but differed only slightly from solutions obtained with a lower service level of ten %. At 70 coils, the optimization of a particular seed yielded an exceptionally high solution with both the default and the tuned solver.

Table 9: Results of seed 100 in optimizations with the tuned and the default solver, 70 coils and a high service level.

	# stringers	# solutions	MIPGap	Tardiness
Tuned solver	67	1	0.970	5
Default solver	65	1	0.985	14

As shown in table 9, these extraordinarily high objective values were the only feasible solution the solvers were able to calculate in time. The high MIPGap of both solutions suggests, that the optimality of these solutions either could not be proved in time or that they are not optimal, with the latter being the more plausible explanation when regarding the high objective values. The most significant difference between these solutions is the tardiness, which is much higher with the default solver.

As previously described, the tuning should have little impact on the solutions themselves. This, combined with the fact, that it was the only feasible solution found, suggests that this more optimal solution, regarding tardiness, is a result of randomness. Even though these solutions could be characterized as outliers, they were still included in figure 3 because they represent feasible solutions and could technically be implemented, despite being probably insensible businesswise.

More exceptions to the downward trend are the solutions yielded at an instance size of 20 coils across all service levels. This is the same instance size that resulted in infeasibility with two seeds, as described at the beginning of the section. The objective values of nearly every seed across all service levels increased when transitioning from ten to 20 coils, and

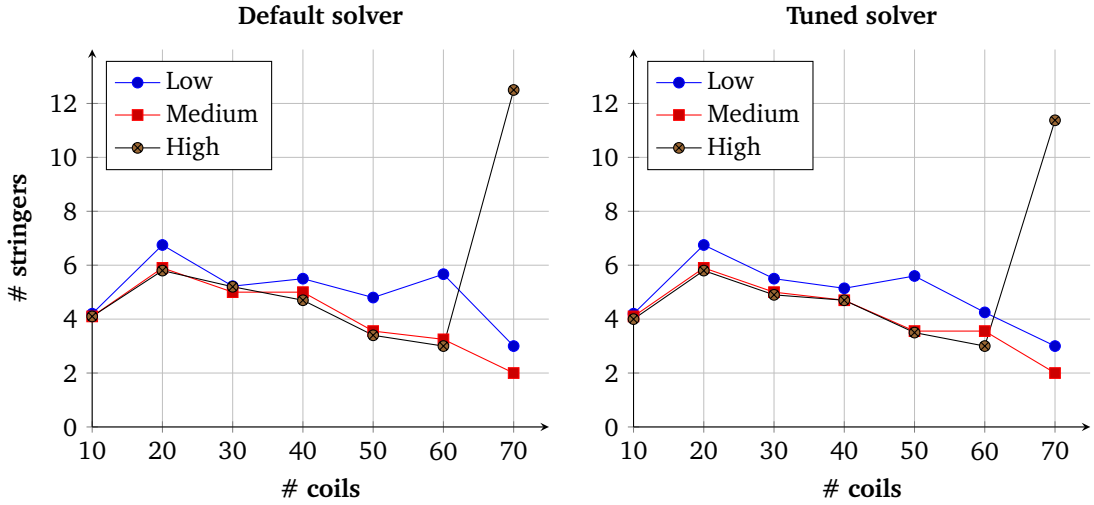


Figure 3: Average number of introduced stringers per $\mathcal{N}$ and service level in optimizations with the default and tuned solver.

therefore it is not a result of an outlier as it was in the case of 70 coils. Additionally, this suggests that this is not a random occurrence but a trend. Hence, it might be beneficial to observe the relative average number of stringers introduced, which are depicted in figure 4.

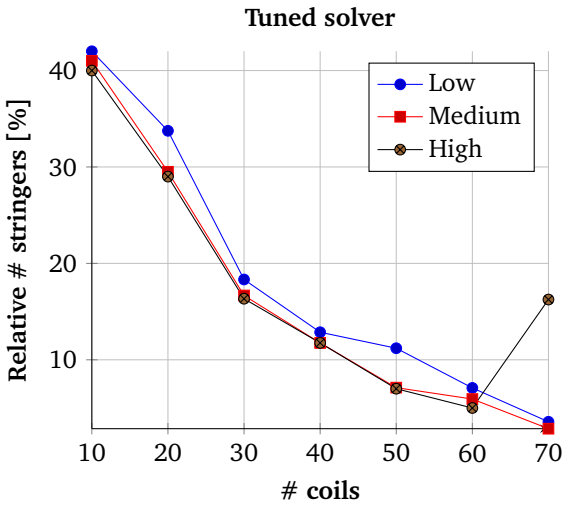


Figure 4: Average number of stringers relative to the instance size of optimizations with the tuned solver in %.

Figure 4 portrays the average number of stringers relative to the instance size of the optimizations. To increase the visualization, only data from the tuned solver is shown. The figure depicts, that the absolute numbers of coils are deceiving in this regard and that in fact, the results from the optimization with 20 coils are in line with the observed downward trend. Now that the existence of a constant downward trend, except for the case with seed 100 mentioned above, is established, a possible explanation can be discussed.

To recapitulate, a stringer has to be introduced between adjacent coils if the difference between the physical or mode-specific or both characteristics of these coils is above the limit specified by the process flexibility PF. The introduction of a stringer can be avoided if the schedule introduces a third coil between these incompatible coils, but this is only possible if the third coil is compatible with the other two coils. An explanation for the high relative number of introduced stringers in optimizations with small instance sizes could therefore be the absence or an insufficient number of these third coils that can bridge the gap between two incompatible coils. By increasing the number of coils, these gaps could then be closed or narrowed by the new coils. The likelihood of this increases with the increasing number of coils, which results in fewer stringers that have to be introduced due to differences in characteristics.

To give a simplified example, in which only the annealing temperature of the coils is considered as compatibility restriction: Three coils have to be scheduled in sequence, with coil one having an annealing temperature of 670 °C, coil two having an annealing temperature of 740 °C and coil three with an annealing temperature of 720 °C. The PF is 40 °C in a symmetric interval and the annealing temperature is only distributed in increments of ten °C, as explained in section 4.1. Thus, coil two and three are compatible with each other, but not with coil one. Since there is no other coil to bridge this gap, a stringer has to be added between them. Now, a fourth coil has been added to the schedule with an annealing temperature of 700 °C. It is therefore compatible with coil three, but not with coils one and two. Since no coil is compatible with coil one, a stringer still has to be added, but the relative number of stringers decreased from 33,3 % to 25 %. If another coil with an annealing temperature of 680 °C or 690 °C would be added, it would close the gap and lead to a stringer-free schedule. If it has a different annealing temperature though, it would still be compatible with one of the remaining coils and therefore, still only one stringer would

be necessary, decreasing the relative number of stringers even further. By adding more and more coils to this schedule, the probability that one of the coils eventually closes this gap increases.

The reason for the increase in the number of absolute stringers from ten to 20 coils could therefore be that most of the existing gaps could not be closed by the new coils and that some coils may have characteristics which were incompatible with both ends of the gap, thus increasing the number of stringers that have to be introduced. To summarize, the probability of a single coil being compatible with the other coils in the schedule is higher for a higher number of coils. Thus, these schedules with a higher number of coils need fewer stringers, since there are fewer gaps and more possibilities to sequence these coils to avoid gaps. In the actual optimizations performed in this study, there are more compatibility restraints as well as due dates to consider, which is why in most cases there are still stringers introduced in the schedule, even with higher numbers of coils. The impact of the restriction by due dates can be observed by comparing the average of the optimizations with a high and low service level because, in fact, most optimizations do not utilize the entire service level. This is portrayed in figure 5.

Since the service level is a parameter to limit the total amount of delayed coils relative to its instance size, the threshold for relative tardiness for a high service level is 20 % and so forth. As mentioned above, it can be observed that on average and with a high service level, at no instance size the full service level of 20 % is leveraged, with the highest average being 19,25 %. The data suggests, that for most cases a service level of 16 % to 17 % would suffice. This is supported by the fact, that optimizations with a service level of ten % always utilize the entire service level or almost all of it. This does not imply that if the service level would be increased to over 20 %, the solver would not leverage it. But it states, that to receive the solutions displayed in figure 3, a service level under 20 % would suffice, at least in most cases. As expected with a low service level, the solver does not allow a delay of coils.

The data also suggests a slight decrease in delays in optimizations with a higher number of coils through the tuning of the solver. But, as previously described, this could be affected by the selectivity effect. When comparing the results of figure 3 and figure 5, it can be observed that even though solutions found with a service level of 20 % always had a higher number of delays, they did not necessarily yield more optimal solutions in terms of stringers than optimizations with a service level of only ten %. In fact, the most significant difference in average stringers introduced between the two parameterizations was only 0,556 stringers, which occurred in the case of an instance size of 60 coils with the default solver and thus could be influenced by the selectivity effect.

The question that could arise is, if not profoundly positively affecting the solutions in terms of introduced stringers, what is the benefit of using a higher service level? This question could be answered by taking a look at figure 6.

The data of higher instance sizes, especially 60 and 70 coils, seems to be strongly affected by the aforementioned selectivity effect and will therefore be disregarded in the following discussion concerning the total optimization time. The top graphs portray, that a higher service level usually leads to a lower average total optimization time, which could save costs caused by the operation of the computer, but that in general, the total optimization time increases profoundly. This trend can be explained by the bottom graphs, which display the extraordinary increase in variables and constraints, and thus the size of the MILP. The stark increase in variables is mostly driven by binary variables, like y_{ijkmn} and x_{ikm}. Since the number of variables and constraints are not influenced by the parameterization of the service level or tuning, they are only displayed by one graph each. When comparing the top with the bottom graphs, a positive correlation can clearly be observed. Therefore, the probable explanation for the increase in total optimization time is the growth of the model the solver is trying to solve.

The size of the model has two different impacts on the total optimization time. The first and most obvious is, that it has to consider a growing number of constraints and variables when trying to find a feasible solution, which increases the time it takes to prove the feasibility of it. But before the model can be solved, it first has to be built, which is the second impact of the model size. The time it took the solver to build the model rose from only 4,5 seconds with ten coils to over 90 seconds with 50 coils, which does not seem to be affected by the service level or by tuning. Therefore, the model should be kept as small as possible to avoid wasting optimization time, which is especially important when certain circumstances severely limit it. What seems to be affected by the service level and by tuning is the solving time, which accounts for the majority of the total optimization time. Tuning can in fact increase the solving time through different parameterizations of $PrePasses$ and $Presolve$, as an example.[54] In the case of this study, the tuning thus seems to increase the success rate of optimizations, especially with greater instance sizes, at the cost of solving time. A higher service level, on the other hand, seems to decrease the solving time, which suggests that the solver can find optimal solutions faster if it is less restricted. The MIPGap also seems to benefit from an increased service level, as depicted in figure 7.

A positive trend between the MIPGap and the instance size can be observed, as well as the trend that optimizations with more scheduling flexibility tend to have a higher grade of proven optimality, which can be explained by the lower solving times. The lower the service level, the more optimizations reach the solving time limit of 720 seconds and thus have to abort the solving process. Hence, the solver has no or less time to prove the optimality of the solution, resulting in a higher MIPGap. This does not necessarily result in a worse solution, as the solutions, in terms of stringers, between a service level of ten % and 20 % only differed slightly, as portrayed in figure 3. Thus, the general upwards trend of

[54] Gurobi, n.d.-b

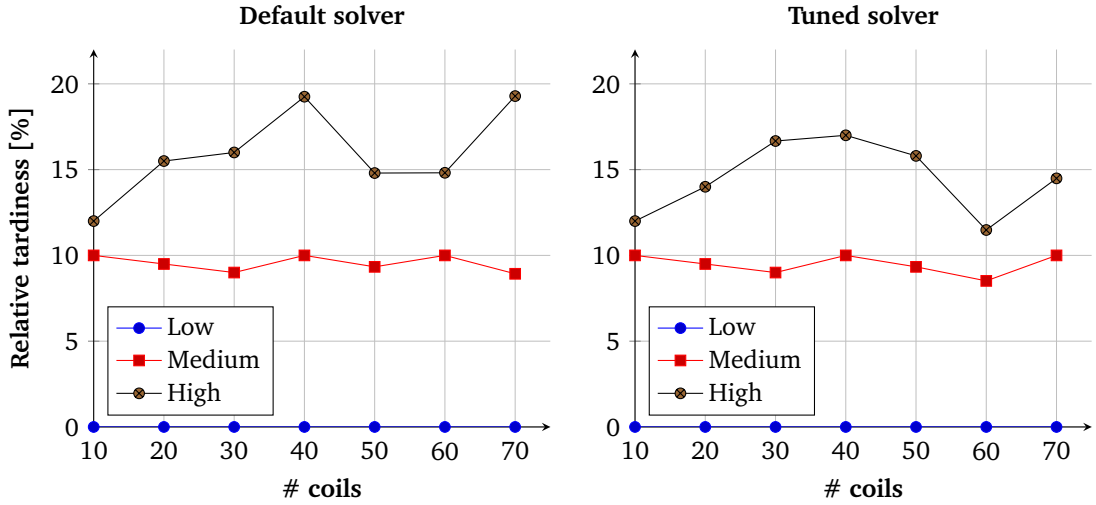


Figure 5: Average tardiness relative to the different instance sizes per service level in optimizations with the default and tuned solver.

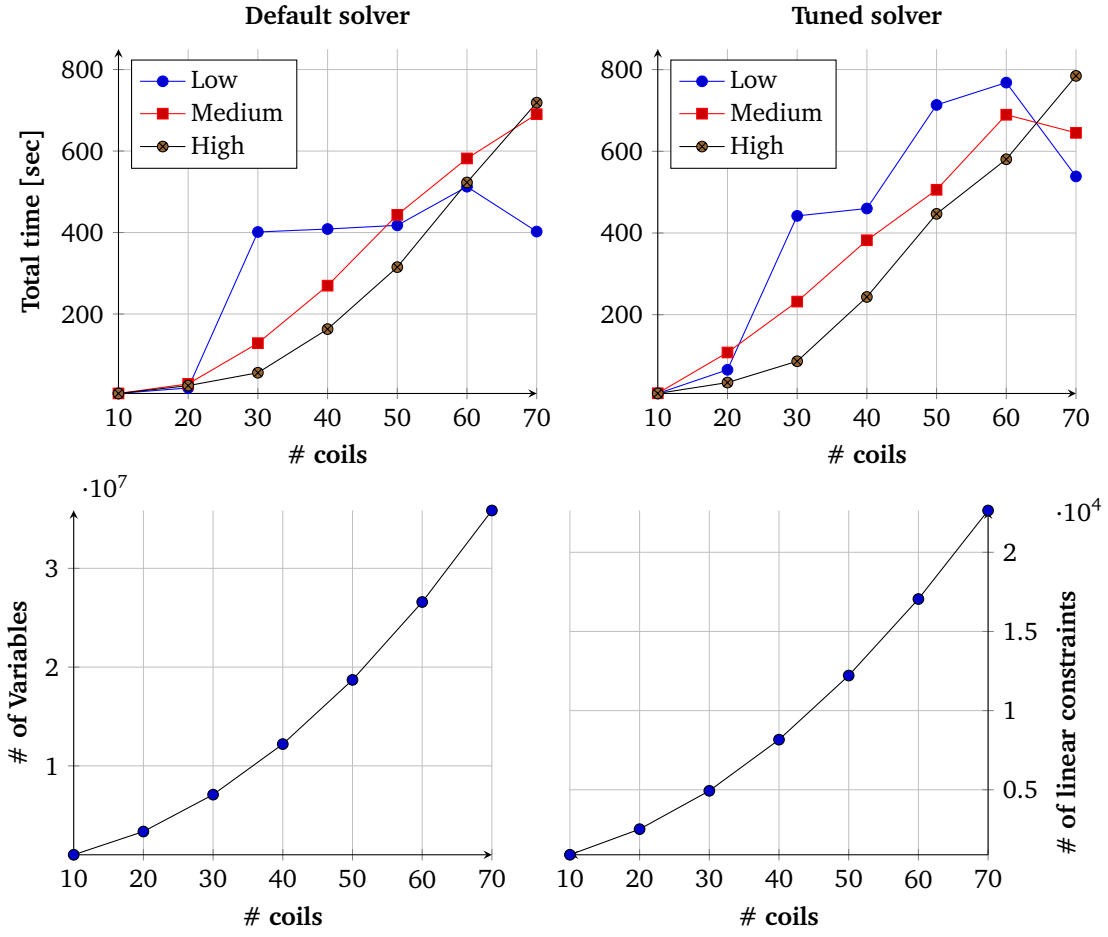


Figure 6: Average total optimization time (top) and number of variables (bottom-left), as well as average number of linear constraints (bottom-right) per service level and instant size in optimizations with the default and tuned solver.

the MIPGap could be a consequence of the increasing solving times of each seed. Hence, it could be advantageous to choose a higher service level if the available optimization time is severely limited or if a higher proven grade of opti-

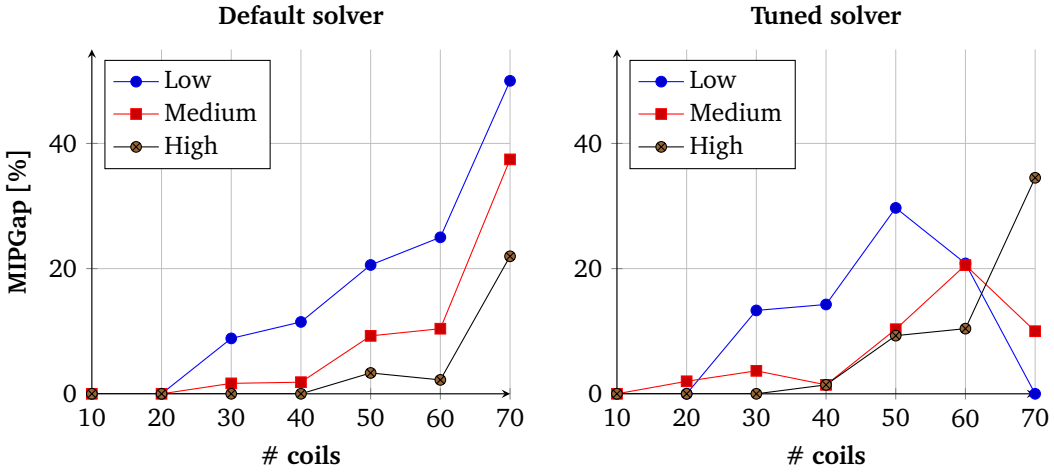


Figure 7: Average MIPGap per instance size and service level with the default and tuned solver in %.

mality of the feasible solution is preferred. The data from the tuned solver seems to be affected by the selectivity effect, but the MIPGaps of smaller instances seem to be similar to that of the default solver. The mentioned increase in solving times is displayed in figure 8 with data from optimizations solved by the default solver due to its better visualization of this trend. All boxplots were created with the median and the highest and lowest value from the ten optimizations, as well as the 25 % and 75 % quartile.

Figure 6 already depicted the increasing average total time. But this data could suggest that this applies to every seed, when in fact it does not. The data in figure 8 portrays that, starting from an instance size of 30, the differences in solving times increase drastically if they are solved with the two lower service levels. This increase can be observed from 40 coils onward, with the highest service level. The higher number of optimizations reaching the time limit combined with the higher average MIPGaps in optimization bound by the lowest service level indicates, that the solver has no time to prove the optimality of the feasible solution after finding it. With a medium service level, fewer optimizations have to be aborted due to the time limit and thus have time to prove the optimality of the feasible solution, which results in a lower average MIPGap. This is also true for the highest service level that displays the fewest number of time limit terminations.

This concludes the first part of the numerical studies. Based on the presented data, an instance size of 40 coils was chosen for the following sections since it is the highest number of coils with a tolerable amount of failed optimizations. In addition, the optimizations were conducted with the tuned solver to increase the number of successful optimizations.

4.2.3. Best- and Worst-case scenario

For the third part, a Best- and Worst-case scenario was derived from the different factors defined by Wegel et al. (2024), which were described in section 4.1. The results will be compared with the results achieved with the tuned solver from the previous section. In addition, the PF and HC were altered because it was suspected that these factors are the most influential among them. A more detailed analysis of the influence of each factor and level on the scheduling is provided by Wegel et al. (2024).

Figure 9 compares the number of stringers introduced between the Best-case scenario with a *Very low* PF and the *Basecase* scenario. The data from the *Very high* and *Basecase* PF is not portrayed, as no stringers are introduced at any point in these scenarios. As previously observed, the number of stringers decreases with higher service levels in the Best-case as well due to the higher scheduling flexibility. Compared with the results from the *Basecase* scenario, the Best-case yields better solutions, in terms of stringers, for every level of PF and for every service level. Even with *Very low* PF, a stringer reduction of over 50 % can be observed. Thus, a lower PF leads to an increase in stringer introduction, but in this case, this effect is probably reduced by the *Very low* HC. Since these effects could therefore counteract each other, the stringer reduction may be a result of the lower UC and SPT. If true, this could also result in fewer delayed coils, which are depicted in figure 10.

By comparing the average tardiness in the Best-case scenario - *Very low* PF with the tardiness from the *Basecase* scenario, only a slight decrease can be observed. This difference increases with the B. PF, but does not with the highest PF. Thus, the *Very low* UC and SPT do not lead to severely less tardiness when HC and PF are parameterized as *Very low*. But the increased scheduling flexibility through more relaxed time constraints could still benefit the optimization, thus decreasing the number of introduced stringers.

Figure 10 also displays the distribution solving times. The data suggests, that the level of PF, in the context of the Best-case scenario, has an immense impact on the solving time. The optimizations with the lowest PF have the highest solving times, even higher than in the *Basecase* scenario. But by increasing the PF to the *Basecase* level, the median solving

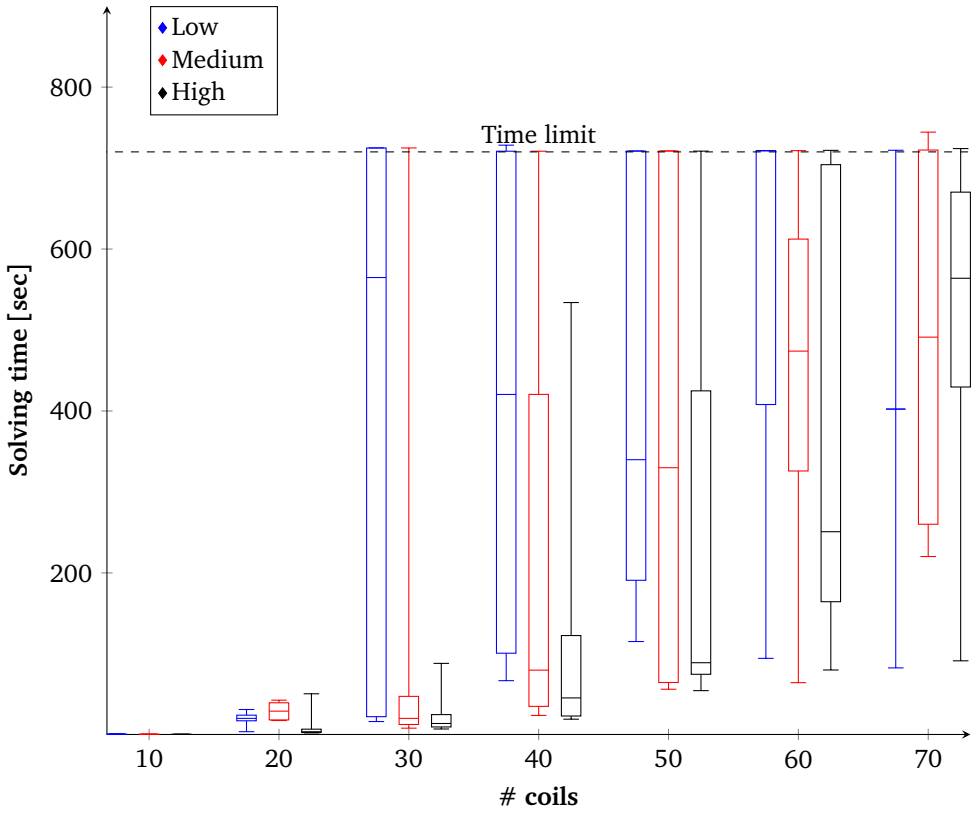


Figure 8: Distribution of solving times in seconds per instance size and service level with the tuned solver.

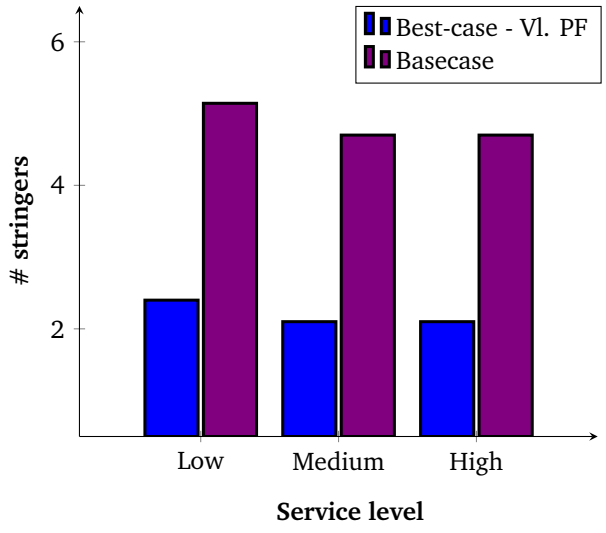


Figure 9: Average number of introduced stringers per service level in the Best-case scenario with *Very low* PF and *Basecase* scenario.

times decrease to a fourth. Further increasing the PF to *Very high*, however, does decrease the solving times only slightly. That the solving times for the two higher investigated PF levels are remarkably lower than in the *Basecase* scenario can be explained by the parameterization of the HC, UC and SPT. As previously described, the *Very low* UC and SPT increase the scheduling flexibility by relaxing the time constraints. The *Very low* HC, on the other hand, increases the general compatibility between different coils, which leads to more feasible, stinger-free processing sequences. All of these factors could result in more solutions being feasible, thus decreasing the time it takes the solver to find an optimal, feasible solution. This could be applied to the lowest PF as well but in reverse. It could suggest that the *Very low* PF drastically limits the number of feasible, stringer-free processing sequences of adjacent coils, which increases the solving time since the use of stringers has to be minimized. In this regard, the other very beneficially parameterized factors cannot offset the *Very low* PF, which could indicate that the PF is the most impactful factor of the four. This statement could be confirmed by the data from the Worst-case scenario.

Figure 11 portrays the impact of the HC on the Worst-case scenario in terms of introduced stringers and failed runs. As previously described, *Very low* HC is beneficial for stringer-free sequences. In this case, the HC seems to impact the average number of stringers more than the other factors, displayed by the fact that the lowest HC in combination with the Worst-case yields better solutions than the *Basecase* scenario. This would contradict the statement, that the PF is the

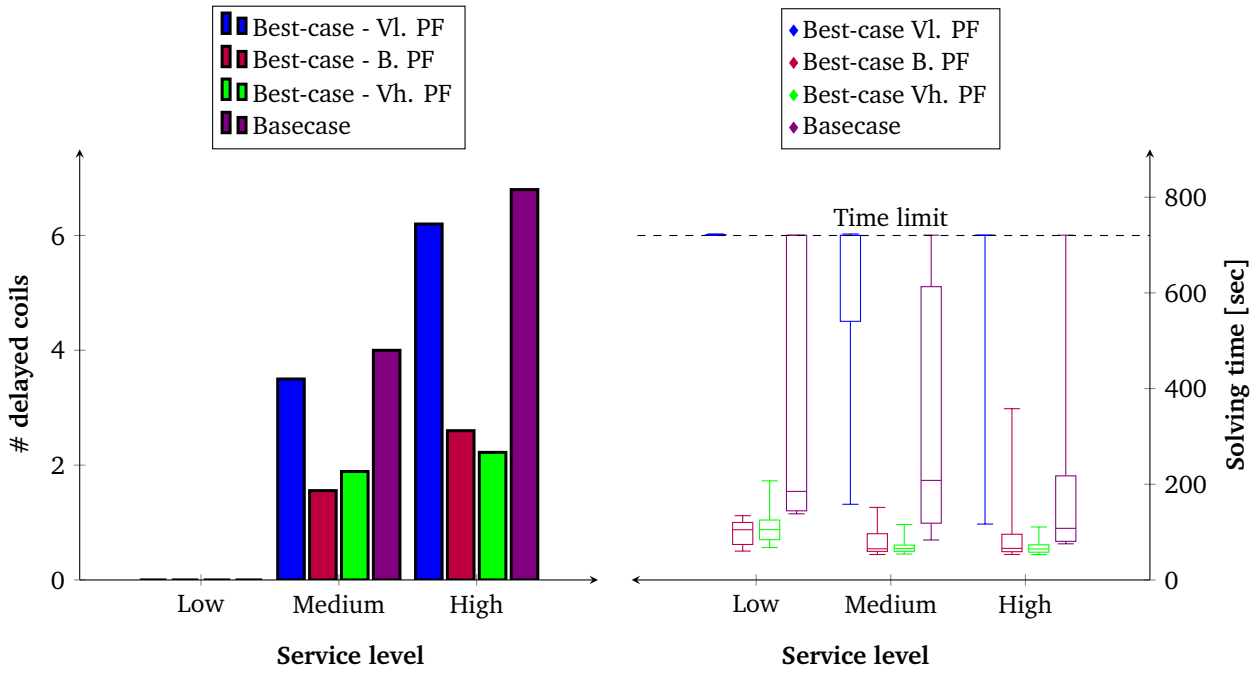


Figure 10: Average number of delayed coils and distribution of solving times in seconds per service level in the Best-case instances and *Basecase* scenario.

most influential factor. But the data from the two other HC parameterizations support this claim. While the optimizations with Worst-case and *Basecase* HC did still yield some but very high solutions, no optimization was able to calculate a feasible solution with the highest HC.

In general, the Worst-case scenario seems to negatively impact the number of successful optimizations. As previously described, the reasons are twofold. While all runs with the lowest service level fail due to infeasibility, optimizations with a higher service level fail because of the solving time limit, with two exceptions with a medium service level and *Very high* HC. This demonstrates the great impact of the Worst-case and especially of the HC on the feasibility of the model. Before finalizing the second part of this numerical study, one last test was conducted to analyze which of the two factors, HC and PF, impact the solution the most. To accomplish this the case in table 10 was defined.

Table 10: Case parameterization to identify the most impactful factor.

Case	HC	UC	PF	SPT
Impact	Very low	Basecase	Very low	Basecase

The figure 12 displays the differences in stringers and delayed coils between the *Basecase* scenario and the Impact case, by subtracting the average number of stringers and delayed coils in the Impact case from the *Basecase* scenario.

It can be concluded, that the PF has a greater impact on these aspects of the solution than the HC. This observation was confirmed by a second Impact case, which used the *Basecase* scenario and *Very high* PF and HC. But it has to be noted that this only applies to the definitions of the different levels of HC and PF by Wegel et al. (2024). Both factors indicate a strong influence on the characteristics of the solution and should therefore be optimized to achieve optimal solutions in real-world applications. After studying the drastic effects of the Best- and Worst-case, different parameterizations of the model will be explored in the following section.

4.2.4. Alteration of processing lines and due dates

First, the number of processing lines in the model was altered. To recapitulate, in the first case, an additional line was added to the model, which can process every coil. In the second case, Removal I, the processing line that can only process wider lines was removed, while the processing line that can only process narrower coils was removed in case three. The line that can process every coil was not removed because it would lead to infeasibility since some coils could not be processed.

Foremost, none of the three cases displayed in figure 13 seems to have a great impact on the number of introduced stringers. As expected, the number of stringers decreased in the addition case and increased in both removal cases, but all these effects were only minor. The reason the addition of one processing line generally reduces the number of stringers is that an incompatible coil or a sequence of coils, that is incompatible with another sequence of coils, can be processed on the additional line, thus reducing the number of stringers by one.

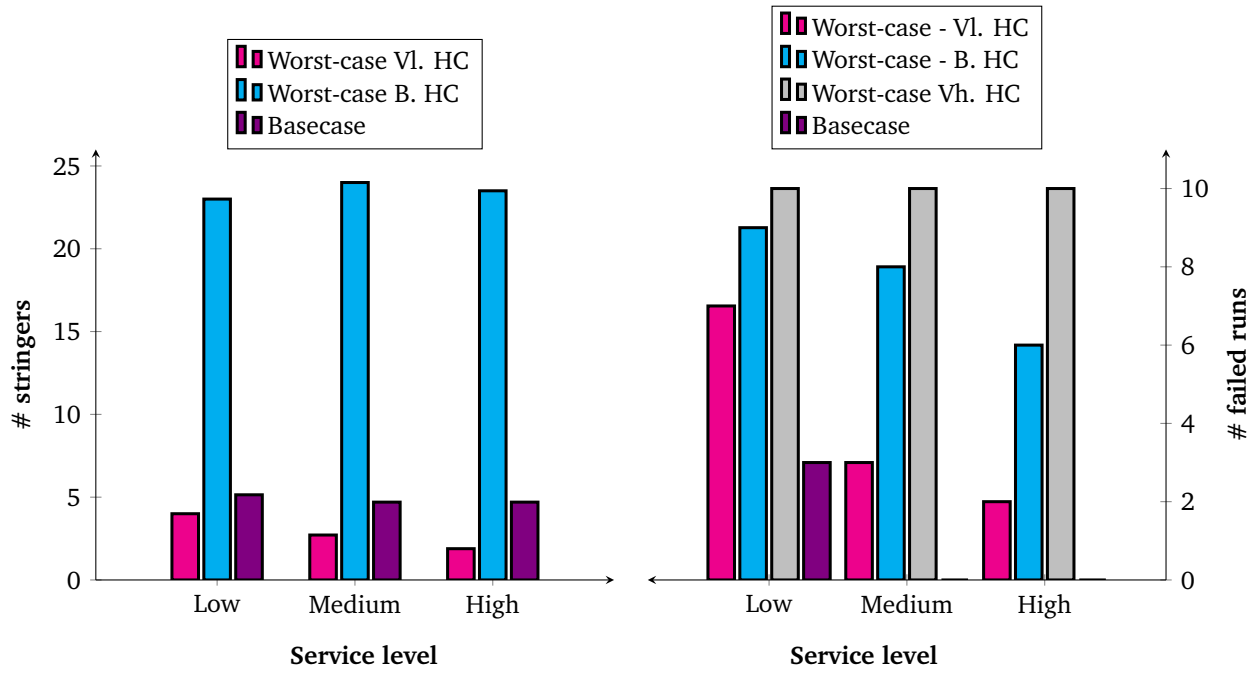


Figure 11: Average number of introduced stringers and failed runs per service level in the Worst-case instances and *Basecase*.

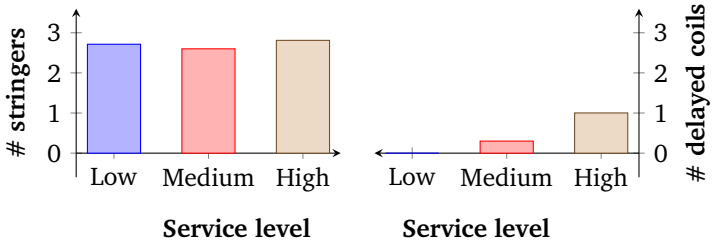


Figure 12: Differences in stringers and delayed coils between the Impact case and *Basecase* per service level.

To give an example, four sequences of coils have to be scheduled, with each coil in a sequence being compatible with every other coil in the same sequence. However, the sequences are incompatible with each other. If these four sequences are processed on one processing line, three stringers would have to be introduced to ensure compatibility. If a new processing line would be added, one of the sequences could be processed on the new line, reducing the number of stringers to two, but nothing more could be done to reduce the number of stringers further.

Therefore, the maximum benefit of adding a processing line, in terms of stringer use, is the reduction of introduced stringers by one. This benefit could increase if the schedule introduced stringers to not violate the service level, but this was not the case for these optimizations.

The reason the data does not show a reduction of the average introduced stringers by one is that two optimizations, that were unsuccessful in the *Basecase* scenario, were successfully solved in the addition case and have a stringer count of six, thus increasing the average number of stringers. In fact, all three cases had only one failed optimization each with a low service level. The explanation for this reduction could be the solving itself since the solver behaves differently even if only one parameter changed.[55] Because the Addition case and the Removal cases have opposite effects, the addition of a processing line increases the number of feasible solutions and the removal decreases them, it is the most likely explanation.

The removal of a processing line influences the number of stringers in the same way as an addition of a line does but in reverse. Thus, if no stringers are introduced to not violate the service level, the maximum increase in stringers is one. This is true for every optimization with a MIPGap of zero %. However, some optimizations with a MIPGap of zero % did not suffer any increase in stringer use, suggesting that they did not utilize the line in the schedule. But some optimizations with a MIPGap of over zero % had an increase of up to three stringers. Hence, the optimizations either had

[55] Miltenberger, 2023b

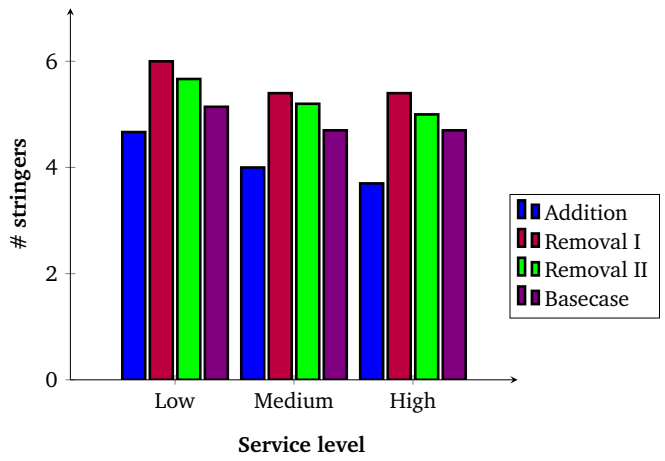


Figure 13: Average number of stringers in the three line-altering cases and *Basecase* scenario per service level.

to introduce stringers to comply with the service level and the solver was unable to prove the optimality of the solution in time or the solver simply could not find the optimal feasible solution until it had to terminate the solving process. This termination happened more often with the addition of one line and is presented in figure 14.

As portrayed, the majority of optimizations with an additional processing line and a low service level had to be terminated due to the time limit. This results in a higher MIP-Gap, as mentioned before. Higher solving times and MIP-Gaps than in the *Basecase* scenario are also observed with higher service levels but are not as severe with the highest one. This overall trend probably results from the increasing model complexity, which could also be observed in figure 6 from the first part of this numerical study. In this case, however, the increase results from the additional line and not from additional coils. The adverse effect can be examined in figure 15.

The data indicates a decreasing solving time through complexity reduction when applying a medium or high service level, but not with a low one. A possible explanation could be, that the restrictions caused by the removal of a processing line in combination with the low service level result in many former solutions being infeasible and thus a longer search for a feasible solution. That Removal II displays lower solving times than Removal I could suggest that the line which can only process wider coils and was removed in Removal I, was utilized to a higher degree than the line that was removed in Removal II. The reason for this is the partially triangular distribution of the coil width due to the *Basecase* level of the HC, which results in more coils being wider than narrower. Hence, the effect of the removal of one line depends on the HC setting.

The last parameter to investigate is tardiness, which indicates to not be severely affected by the addition or removal of a processing line. That the number of delays only complies with the service level and does not have to be minimized could explain why tardiness did not decrease with an additional line. Additionally, no increase in tardiness suggests that most optimizations in Removal I and Removal II had enough scheduling flexibility to find feasible solutions with similar tardiness, even with only two lines.

In contrast, tardiness was the most important factor in the next part of the numerical study, in which the time horizon for the due dates was moved closer to the beginning of the schedule. These earlier due dates are expected to have a certain effect on the number of stringers and delayed coils. With earlier due dates and an imposed service level, the schedule has to implement more stringers to not exceed the allowed tardiness. This further reduces the scheduling flexibility due to the SPT and thus increases the difficulty of finishing the processing of other coils before their due date. Therefore, an increased number of failed optimizations and a higher stringer and tardiness count in successful runs were expected.

The data displayed in figure 16 shows an increasing number of failed optimizations, which aligns with the expectations. In case 100 %, most optimizations are terminated and do not fail because of infeasibility, suggesting that they could be solved with a higher time limit. But this changes with case 50 %, in which no optimization with a low service level is feasible, while most unsuccessful optimizations with a higher service level are terminated. By reducing the values of a and b by 75 %, almost all optimizations with a low and medium service level are infeasible, but only one optimization with a high service level. Therefore, provided with a high service level and a higher time limit, most optimizations could still yield feasible solutions. However, the necessary solving time and the quality of the resulting solution cannot be determined without further studies.

The second expectation was a rise in the number of introduced stringers. Due to the very low number of solutions, the following interpretations could be heavily influenced by the selectivity effect. Because of this, only the most important information will be presented in table 11 and briefly discussed. If a case is not included in a service level, then no optimizations for that case and service level yield a feasible solution.

As indicated by the data, the number of introduced stringers increases through the shift of the due dates. In addition, almost all optimizations completely utilize the service level. Both of these observations represent an increase in comparison to the *Basecase* scenario and align with the expected results, thus concluding the third part of the numerical study.

4.2.5. Numerical studies on the extended models

The last part of the numerical study consists of several tests conducted on the extended models described in sections 3.4.2 and 3.4.3. Both additionally incorporate the release dates of coils from section 3.4.1. The first extended model minimizes stringers and absolute tardiness. First, the average number of stringers and delayed coils in the

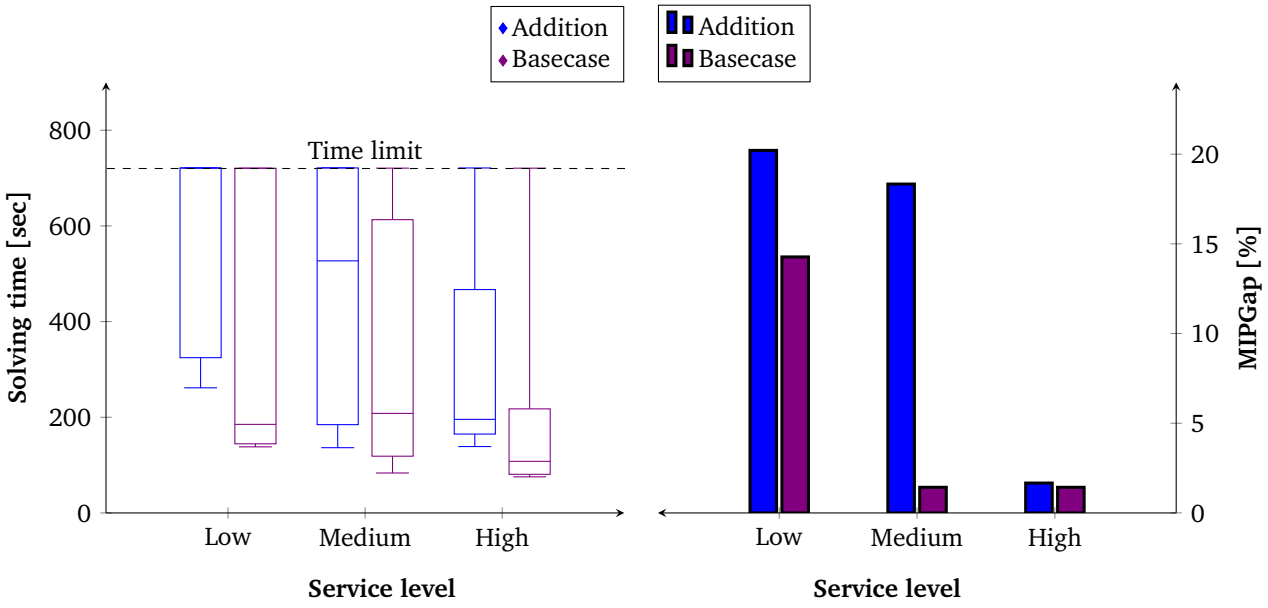


Figure 14: Distribution of solving times in seconds and average MIPGaps in % per service level in the Addition case and *Basecase* scenario.

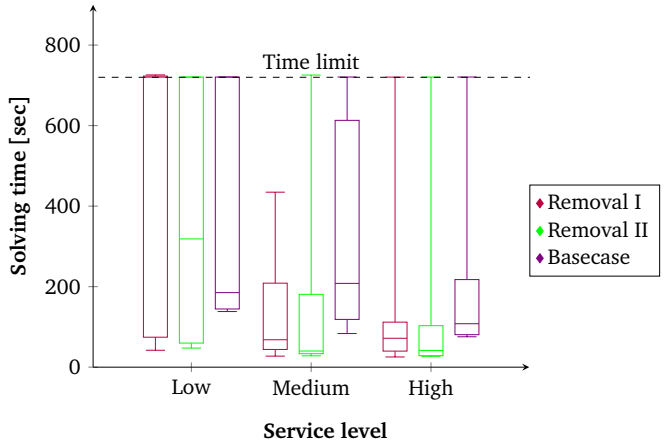


Figure 15: Distribution of solving times in seconds per service level in the Removal I and II cases, as well as in the *Basecase* scenario.

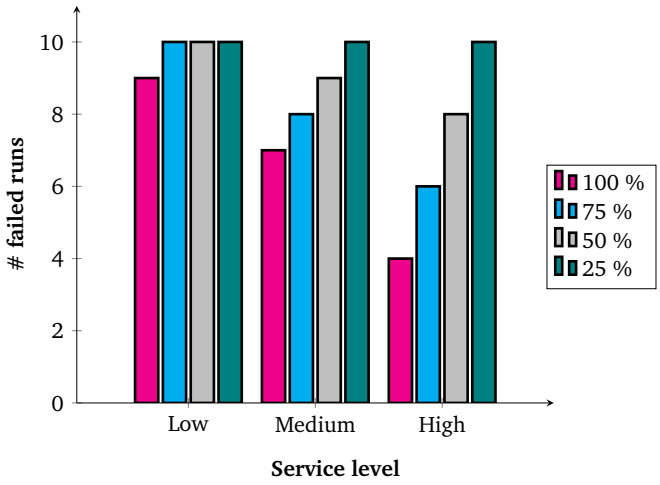


Figure 16: Number of failed runs per service level in the four different due date altering cases.

Basecase scenario with *Very low* and *Very high* UC will be examined.

Figure 17 depicts an expected trend. The number of delayed coils is the highest if it has only a weight of 20 % of the objective value, which has to be minimized. Additionally, the number of stringers is the lowest at this point, since it has a weight of 80 %. Thus, the optimal solutions seem to delay some coils to minimize the number of stringers, since it has a much higher weight. This trend changes, with tardiness being reduced at the cost of stringers with higher weights of tardiness. As portrayed with *Very low* UC, the number of delayed coils is, even at its highest point, very low. Thus, the trade-off between stringers and tardiness cannot be observed as well as with the *Very high* UC on the right side of figure 17.

In this scenario, an almost one-to-one trade-off between the number of delayed coils and stringers can be observed with the increasing weight of the former. The mechanism behind the trade-off is similar to the one discussed in the third part of the study, regarding the addition and removal of processing lines. If the tardiness has almost negligible weight, the coils are sequenced in a way that the number of stringers is minimized with no regard to the number of delayed coils. But if the tardiness weight increases, the objective value may benefit from introducing stringers to finish the processing of

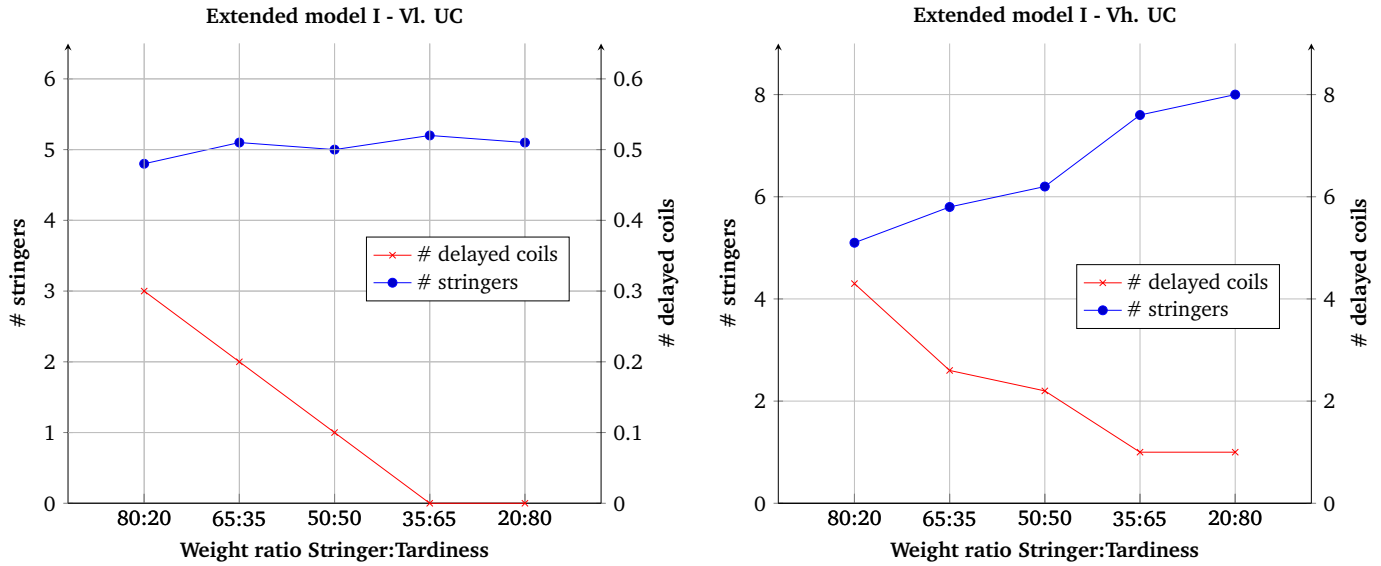


Figure 17: Average numbers of delayed coils and stringers with different UCs and weights in the first extended model.

Table 11: Stringer use and tardiness in the cases 100 %, 75 % and 50 %.

Service level	Case	# stringers	# delayed coils
Low			
	100 %	6	0
Medium			
	100 %	6.667	4
	75 %	7	4
	50 %	7	4
High			
	100 %	5	7.83
	75 %	5.75	8
	50 %	7	8

coils before their due date, thus decreasing the number of delayed coils at the cost of stringers. By further increasing the tardiness weight, more stringers are introduced to comply with due dates and therefore to minimize the objective value. With earlier due dates, this trend is better to observe since it is harder to comply with them, which increases the number of stringers that have to be introduced to decrease the tardiness by one unit and vice versa.

Due to the different optimization objectives, the comparability between the extended model and the base model is limited. But the results of the extended model with a tardiness weight of 20 % are similar to the results yielded with a medium service level in the base model, presented in figure 3. The difference, however, is the UC, which is higher in the extended model and thus highlights its performance to minimize stringer use and tardiness at the same time. But this high performance comes with the cost of increased solving time, portrayed in figure 18.

Generally, there is a wide disparity in solving time between different optimizations in the extended model, regardless of the weight ratio, with most median solving times being higher than that of the *Basecase* scenario with a low service level.

The highest median solving times and terminations in the extended model can be observed with the two higher stringer weights of 80 % and 65 %. The median solving times of the other instances are significantly lower, which could suggest that the minimization of stringers is more complex than that of delayed coils. But this cannot explain, why the median of the solving time increased by increasing the tardiness weight further to 65 %. This effect mostly stems from one optimization, which suffered an increase of 510 seconds in solving time from 210 seconds to 720 seconds. Without this run, the median would be 283 seconds, which is similar to the instances with 50 % and 80 % tardiness weight and could indicate that this optimization is an outlier.

The solving times of the extended model with a *Very high* UC are not depicted, since all of them had to be terminated due to the time limit. Nonetheless, every optimization yielded a feasible solution, which is a drastic increase over the number of failed runs in case 100 % of the third part of the numerical study, displayed in figure 16. Since every result with the extended model and a *Very high* UC also had a low relative tardiness and could therefore comply with the medium and high service levels implemented in the base model at almost every weight ratio, it suggests that it could be more performant to minimize both stringers and tardiness than to implement a service limit and only minimize

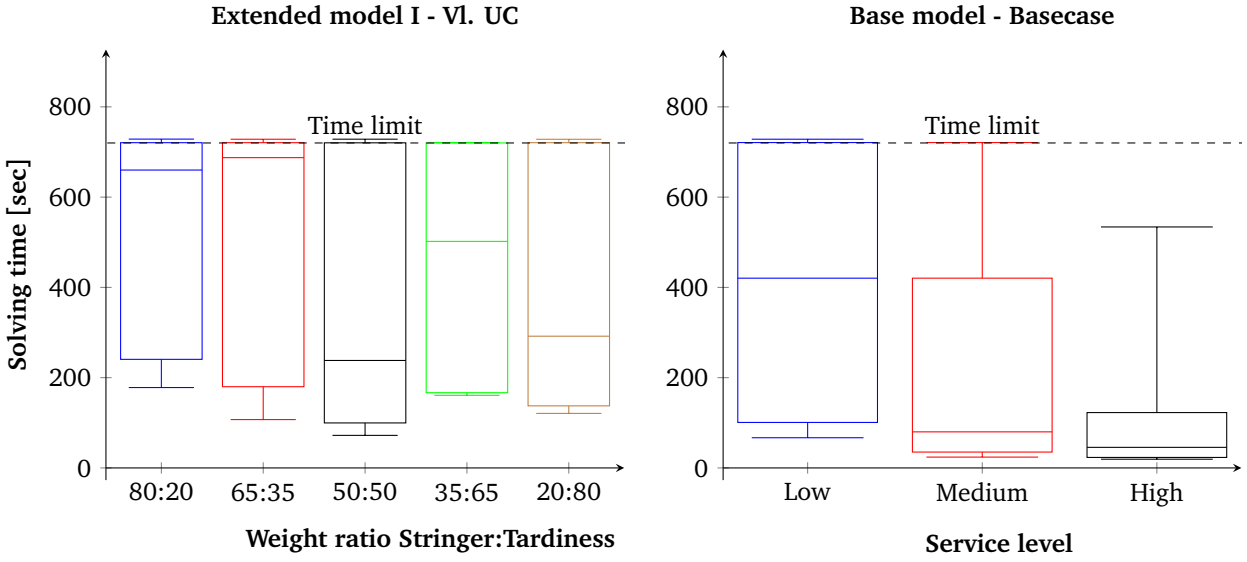


Figure 18: Distribution of solving times in seconds per service level in the base model and per weight ratio in the first extended model.

stringers.

Figure 19 displays a similar trend as in the first extended model, but in this case, the trade-off happens between earliness, tardiness, and stringers. The number of the latter is far greater than in the base and the other extended model. This could be explained by the high MIPGaps resulting from the fact that every optimization had to be terminated due to the time limit.

Therefore, it is possible that the displayed stringer counts are far above the optimal number of stringers. But it has to be noted that every optimization was successful, which is an improvement from the base model. Since the first extended model also had fewer failed runs, the service level might be too restricting, especially in its lower specification. The service level is the most likely cause due to it being the only part that was removed in both extended models.

Due to the lack of reference points, the exact values of earliness and tardiness cannot be compared with other studies. It can be observed, that by increasing the tardiness weight the tardiness can be minimized drastically, but by changing the earliness weight, the earliness only differs slightly. This could result from the different mechanisms of how to minimize tardiness and earliness.

Tardiness can be reduced by finishing the processing of each coil before its due date, which might influence the coil sequencing and thus the stringer use, as previously described. In contrast, to minimize earliness the schedule has to be prolonged, which can be achieved with two methods. The first method would be to introduce stringers in the sequence, even if they are not necessary since they also have a SPT. The second method is to choose longer-lasting processing modes for the coils, which could increase the stringer use as well since the compatibility of these modes would not be as important as it was before. Since stringers still have to be minimized as well, it is more likely that the second method is used to prolong the schedule and that the additional introduction of stringers is therefore only a result of the increasing incompatibility between the processing modes.

Regardless of the reason, an increase in stringer use can be observed with increasing earliness weights, as well as a higher overall stringer use with higher due date deviation and thus earliness weights, which suggests that one of the proposed methods is being utilized. It has to be noted that the change in stringer use is a direct result of the changing earliness and tardiness weights since the stringer and due date deviation weights are fixed throughout the instance. To further prove the earliness minimization mechanism, the optimizations of the instance with a due date deviation weight of 65 % were repeated, but with a $Very\ high$ UC.

Figure 20 portrays a similar stringer use but a significantly lower earliness and therefore supports the proposed mechanism. This is because, as observed in table 11 and figure 17, earlier due dates result in higher stringer use due to the previously described mechanism. But the stringer use did not increase by increasing the UC, which could be a result of less utilization of the proposed earliness minimization mechanism, as this would result in fewer stringers. Additionally, the reduction of stringer use through the reduction of the earliness weight did decrease slightly, but not as much as expected. Therefore, more tests need to be conducted, including alterations of the SPT or the data has to be analyzed regarding the used processing modes and introduced stringers to prove the utilization of the mechanism.

Another observation is the reduction of stringers, earliness and tardiness with certain weight ratios. Since this only occurred with more equal earliness and tardiness weights of 35 % and 50 %, it could indicate that the minimization of multiple objectives benefits from more equal weights between them. But since it did not happen in the instance with a stringer weight of 35 % with *Basecase* UC in figure 19, it

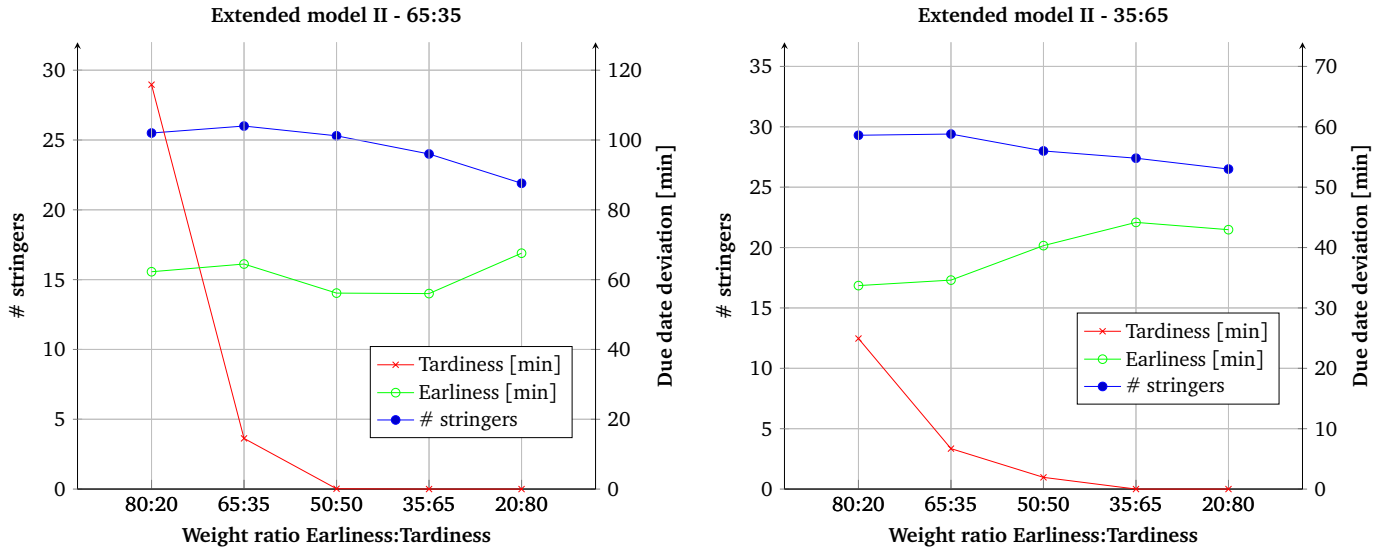


Figure 19: Trade-off between earliness and tardiness in minutes and its effect on the stringer use with different stringer and due date deviation weights in the second extended model.

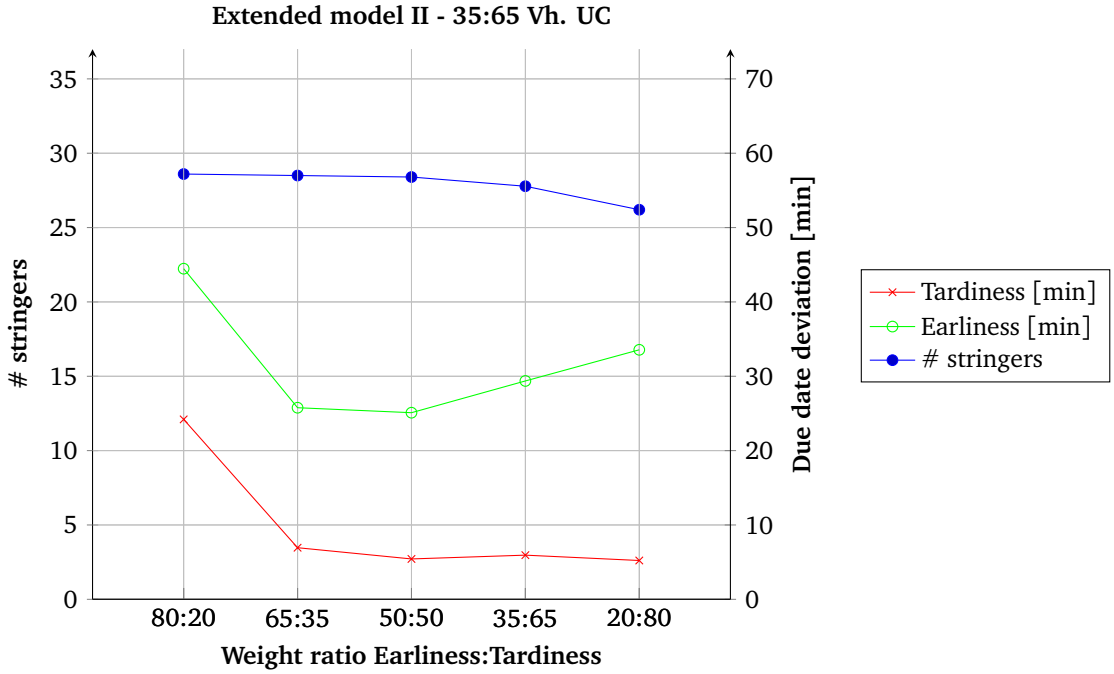


Figure 20: Trade-off between earliness and tardiness in minutes and its effect on the stringer use with a stringer and due date deviation weight of 35 % and 65 %, respectively, in the second extended model with a *Very high* UC.

could also be a random occurrence. This concludes the numerical studies, which will be followed by a brief summary of the results and their interpretations, as well as an outlook for future research and managerial insights.

5. Concluding remarks

5.1. Summary

Section 4.2 presented and discussed different aspects of solutions yielded by the Gurobi solver, as well as parameters that affected the scheduling of parallel heterogeneous CALs. This section will summarize the most important findings of this numerical study.

As observed in figure 2, the data density decreased with increasing instance sizes due to a surge of optimizations that had to be terminated without a feasible solution because of the implemented time limit. This diminished the ability to determine trends that involve higher instance sizes due to the mentioned selectivity effect, which affected the validity of the data. The tuning of the solver decreased the number of unsuccessful optimizations by almost 20 % while having no effect on the number of introduced stringers or delayed coils. It was, however, observed in figure 4, that the number of stringers per coil negatively correlates with the number of coils, which was explained by the higher probability of compatibility of coils with a greater instance size. Additionally, higher scheduling flexibility through higher service levels enabled solutions with fewer stringers, while also decreasing their total optimization time and MIPGap. Figure 6 displayed the surge of the total optimization time due to the increasing size and complexity of the model, as well as a slight increase through the tuning of the solver. A positive correlation between the MIPGap and the instance size was portrayed in 7, which was explained by the increasing number of terminations due to the time limit. It was also argued that a higher MIPGap does not necessarily indicate a less optimal solution, but primarily a deficiency in solving time. This deficiency was depicted in figure 8, as well as the substantial differences between different seeds, in terms of solving time. For the rest of the numerical study, the instance size of 40 coils in combination with the tuned solver was chosen due to their optimal trade-off between data density and the number of coils.

The results from the Best- and Worst-cases utilized in the second part of the numerical study, displayed in figures 9 to 12, were as expected. It was observed, that the parameterizations had a profound impact on the number of successful runs, stringer use and tardiness, as well as on the solving time. While optimizations with the Best-case and $Very\ high$ PF were able to schedule all coils without any stringers, none of the optimizations with the Worst-case and $Very\ high$ HC yielded any feasible solution. By altering the PF and HC respectively, the impacts could be mitigated, and it was found that both factors greatly influence the results of optimizations. An additional test proved that the impact of the HC on the solution is greater than that of the PF, at least for the used definitions of these factors.

In contrast, the impact of altering the number of processing lines on the stringer use was minimal. It was argued, that if no stringers have to be added to comply with the service level, the maximum absolute change in stringer use is one. In fact, the number of introduced stringers did not increase or decrease further for most optimizations through the removal or addition of a processing line, respectively, as depicted in figure 13. But it was observed in figure 14 that it did severely impact the solving times due to the reduction or extension of the model, respectively.

Another severe influence was observed in figure 16 by drastically increasing the UC. A negative correlation between the earliness of due dates and the number of introduced stringers and delayed coils was found, but most optimizations were not successful. The main reason was the drastically increasing share of infeasible models. But the data also portrayed the fact, that most models could still yield feasible solutions even with extremely early due dates, if they are provided with a higher service level and time limit.

By removing the service level and integrating the tardiness in the objective function, the first extended model was able to yield results similar to the base model with a medium service level. However, the former model had a higher UC, as portrayed in figure 17. Additionally, an almost one-to-one trade-off between the number of delayed coils and stringers was observed with $Very\ high$ UC. Figure 18 also indicated that the minimization of stringer use might be more complex than the minimization of tardiness through higher solving times with higher stringer weights.

With the introduction of earliness into the minimization objective, the trade-off between stringer use, tardiness, and earliness was displayed in figure 19 with the second extended model. The observations suggested that the minimization of earliness is more difficult than the minimization of tardiness and that it also correlates with a higher stringer use.

Furthermore, both extended models had significantly higher solving times than the base model, suggesting a higher complexity of the optimization process. In contrast, however, they also had a lower failure rate of zero %, indicating a negative impact of the service level on the optimizations conducted with the base model.

5.2. Outlook, further research and managerial insights

Each of the 1180 optimizations was conducted on a regular household computer. With more sophisticated hardware, much greater instance sizes could be solved while it would also increase the performance of each optimization already studied in this work. This could further be improved by the implementation of heuristics, as described in section 2.2 and allow for a higher number of runs per parameterization, thus improving the validity of the data. Additionally, a test could be conducted which combines the addition of processing lines with extremely early due dates. This should decrease the processing load per line and should therefore increase the feasibility of the optimizations and the optimality of the observed solutions. This test could also be conducted with the second extended model since it should benefit the earliness minimization and thus the stringer use. Moreover, the necessary solving time and quality of the resulting solution of optimizations with case 25 % from the third part of the numerical study could be determined.

But even without better hardware, the solver could be more fine-tuned to increase performance further. Moreover, the studies conducted on the base model during this numerical study could be applied to the extended models to investigate, if the results differ from each other and if so, why. Further extensions of the model are also possible. As an example, the release dates or due dates could be replaced with the entire upstream or downstream process, respectively. This would be one step further to scheduling the whole steel processing complex. Another study could try to investigate the

cause of the high differences between different seeds. Possible explanations could be either the different distributions of coil characteristics, due dates etc. or the randomness of the Gurobi solver if provided with different parameterizations.[56] Furthermore, the processing load of each processing line in different scenarios could be studied, which could have some implications for real-world applications. But some managerial insights can already be derived from the observations made in this study.

The impacts of the Worst- and Best-case scenarios should be considered when scheduling CALs. It is difficult to establish the Best-case in real-world scenarios because most factors are determined by the customer order and the manufacturer may not be able to influence them. But the Best-case should be established to the highest degree possible to minimize the incurring costs. Due to the decreasing relative number of stringers, it could also be advantageous to include as many coils in a schedule as possible, but it is not clear how this trend behaves with higher coil numbers. In contrast, the model building time increases disproportionately with the addition of more coils, which could pose a problem if optimization time is limited. Thus, it could be beneficial to organize all coils in smaller, homogeneous batches and put them in an order that minimizes the differences between each batch. Afterwards, each batch could be optimized individually, with the last coil on each line of the preceding batch posing as the first coil on that particular line for the succeeding batch. This could increase stringer use, but would also decrease the total optimization time. Thus, the costs of the additional optimization time would have to outweigh the possible costs caused by the introduction of additional stringers, which could be the case in situations with severely limited available computation time or low stringer cost, or both.

Furthermore, the data indicates that by minimizing both tardiness and stringer use, more optimal schedules could be found, which leads to the final managerial insights. The real-world applicability of the base model may be limited due to the consideration of only the number of delayed coils. While it is important how many coils are delayed, it is also important how long they are delayed. Since the longevity of the delay is not to be minimized in the base model, some coils could be delayed by an, for the customer, unacceptable amount of time. Hence, the second extended model may be more suitable for the scheduling of CALs. But it is also only a deterministic model. A deterministic model must be used complementarily with a dynamic model to account for random events since these naturally occur during operations. In addition, a CFM should be used to improve planning and efficiency over several periods, thus minimizing operational costs over the long term.[57]

Hence, the model and its different variants hold many possibilities for further research, which could result in more managerial insights as well as contribute to the development of a more refined model that could be implemented by a manufacturer to optimize the annealing process on continuous annealing lines.

[56] Miltenberger, 2023b

[57] Iannino et al., 2021, p. 620-630

References

Balaji, P. G., & Srinivasan, D. (2010). An introduction to multi-agent systems. *Studies in Computational Intelligence*, *310*. https://doi.org/10.1007/978-3-642-14435-6

Besson, T. (1998). *Simulation modeling as a tool for assessing the impact of inventory control and scheduling policies in the manufacturing of specialty steel*. Retrieved May 19, 2023, from http://hdl.handle.net/1721.1/50405

Commodity-Inside. (2019). *Flat Steel Market*. Retrieved June 3, 2023, from https://commodityinside.com/flat-steel-market/#:~:text=Cold%5C%20rolled%5C%20steel%5C%20is%5C%20used,which%5C%20require%5C%20good%5C%20surface%5C%20finishing.

Cowling, P., Ouelhadj, D., & Petrovic, S. (2004). Dynamic Scheduling of Steel Casting and Milling using Multi-agents. *Production Planning and Control*, *15*. https://doi.org/10.1080/09537280410001662466

de Harder, H. (n.d.). *Taking Your Optimization Skills to the Next Level*. Retrieved June 8, 2023, from https://towardsdatascience.com/taking-your-optimization-skills-to-the-next-level-de47a9c51167

Dong, Z., Wang, X., & Tang, L. (2021). Color-Coating Scheduling With a Multiobjective Evolutionary Algorithm Based on Decomposition and Dynamic Local Search. *IEEE Transactions on Automation Science and Engineering*, *18*(4), 1590–1601. https://doi.org/10.1109/TASE.2020.3011428

Durlinger, P. (2015). *Inventory and holding costs*. https://doi.org/10.13140/RG.2.1.3478.7684

Gao, C., Tang, L., & Wang, Y. (2008). Model and scheduling of a continuous galvanizing line. *2008 IEEE International Conference on Service Operations and Logistics, and Informatics*, *2*, 1829–1834. https://doi.org/10.1109/SOLI.2008.4682827

Gurobi. (n.d.-a). *GUROBI OPTIMIZER REFERENCE MANUAL*. Retrieved June 6, 2023, from https://www.gurobi.com/wp-content/plugins/hd_documentations/documentation/9.0/refman.pdf

Gurobi. (n.d.-b). *MIP Models*. Retrieved June 6, 2023, from https://www.gurobi.com/documentation/9.5/refman/mip_models.html

Gurobi. (n.d.-c). *Mixed Integer Programming Basics*. Retrieved June 6, 2023, from https://www.gurobi.com/resources/mixed-integer-programming-mip-a-primer-on-the-basics/

Harjunkoski, I., & Grossmann, I. (2001). A decomposition approach for the scheduling of steel plant production. *Computers & Chemical Engineering*, *25*, 1647–1660. https://doi.org/10.1016/S0098-1354(01)00729-3

Iannino, V., Colla, V., Maddaloni, A., Brandenburger, J., Rajabi, A., Wolff, A., Ordieres, J., Gutierrez, M., Sirovnik, E., Mueller, D., & Schirm, C. (2021). Improving the Flexibility of Production Scheduling in Flat Steel Production Through Standard and AI-Based Approaches: Challenges and Perspectives. In I. Maglogiannis, J. Macintyre, & L. Iliadis (Eds.), *Artificial Intelligence Applications and Innovations* (pp. 619–632). Springer International Publishing.

Jiyuan-Shenzhou-Industry. (n.d.). *Payoff reel*. Retrieved May 21, 2023, from http://www.jyszmachinery.com/product/Payoff-reel_1.html

Joint-Research-Centre. (2022). *EU climate targets: how to decarbonise the steel industry*. Retrieved May 20, 2023, from https://joint-research-centre.ec.europa.eu/jrc-news-and-updates/eu-climate-targets-how-decarbonise-steel-industry-2022-06-15_en

Karakostas, P., Sifaleras, A., & Georgiadis, M. (2019). A general variable neighborhood search-based solution approach for the location-inventory-routing problem with distribution outsourcing. *Computers & Chemical Engineering*, *126*, 263–279. https://doi.org/10.1016/j.compchemeng.2019.04.015

Karakostas, P., Sifaleras, A., & Georgiadis, M. (2020). Adaptive variable neighborhood search solution methods for the fleet size and mix pollution location-inventory-routing problem. *Expert Systems with Applications*, *153*, 113444. https://doi.org/10.1016/j.eswa.2020.113444

Krukowska, E. (2021). *Europe CO2 Prices May Rise More Than 50% by 2030, EU Draft Shows*. Retrieved June 4, 2023, from https://www.bloomberg.com/news/articles/2021-06-29/europe-co2-prices-may-rise-more-than-50-by-2030-eu-draft-shows#xj4y7vzkg

Li, T., Meng, Y., & Tang, L. (2023). Scheduling of Continuous Annealing With a Multi-Objective Differential Evolution Algorithm Based on Deep Reinforcement Learning. *IEEE Transactions on Automation Science and Engineering*, 1–14. https://doi.org/10.1109/TASE.2023.3244331

Miltenberger, M. (2023a). *What is the MIPGap?* Retrieved June 5, 2023, from https://support.gurobi.com/hc/en-us/articles/8265539575953-What-is-the-MIPGap-

Miltenberger, M. (2023b). *Why does Gurobi perform differently on different machines?* Retrieved June 4, 2023, from https://support.gurobi.com/hc/en-us/articles/360045849232-Why-does-Gurobi-perform-differently-on-different-machines-

Mirjalili, S., & Gandomi, A. H. (2023). Chapter 21 - Employment of bio-inspired algorithms in the field of antenna array optimization: A review. In *Comprehensive Metaheuristics* (pp. 393–406). Academic Press. https://doi.org/10.1016/B978-0-323-91781-0.00021-1

Mujawar, S., Huang, S., & Nagi, R. (2012). Scheduling to minimize stringer utilization for continuous annealing operations. *Omega*, *40*(4), 437–444. https://doi.org/10.1016/j.omega.2011.08.007

Mujawar, S., Huang, S., Patel, D., & Nagi, R. (2004). Scheduling to Minimize Stringer Utilization in Metal Industries. *Institute of Industrial and Systems Engineers (IISE)*, *126*, 1–7.

Ouelhadj, D., Petrovic, S., Cowling, P., & Meisels, A. (2004). Inter-agent cooperation and communication for agent-based robust dynamic scheduling in steel production. *Advanced Engineering Informatics*, *18*, 161–172. https://doi.org/10.1016/j.aei.2004.10.003

Pan, Q.-K., Chen, Q.-d., Meng, T., Wang, B., & Gao, L. (2017). A mathematical model and two-stage heuristic for hot rolling scheduling in compact strip production. *Applied Mathematical Modelling*, *48*, 516–533. https://doi.org/10.1016/j.apm.2017.03.067

Pan, Q.-K., Gao, L., & Wang, L. (2019). A multi-objective hot-rolling scheduling problem in the compact strip production. *Applied Mathematical Modelling*, *73*, 327–348. https://doi.org/10.1016/j.apm.2019.04.006

Potts, C., & Strusevich, V. (2009). Fifty years of scheduling: A survey of milestones. *Journal of the Operational Research Society*, *60*, S41–S68. https://doi.org/10.1057/jors.2009.2

Sarna, S. K. (2013). *Annealing of Cold Rolled Steel*. Retrieved June 3, 2023, from https://www.ispatguru.com/annealing-of-cold-rolled-steel/

Sarna, S. K. (2014). *Colour Coating of Steels*. Retrieved June 3, 2023, from https://www.ispatguru.com/colour-coating-of-steels/

Steel-Warehouse. (n.d.). *Annealing Cold Rolled Steel*. Retrieved June 2, 2023, from https://www.steelwarehouse.com/annealing/#:~:text=Certain%5C%20hot%5C%20roll%5C%20grades%5C%20can,a%5C%20residual%5C%20stress%5C%2Dfree%5C%20product.

Takurou, M., Tatsuo, T., & Masayasu, O. (2016). Plant Optimal Control System for No. 5 Continuous Annealing Line (CAL) at West Japan Works (Fukuyama), JFE Steel. *JFE Technical Report*. Retrieved May 20, 2023, from https://www.jfe-steel.co.jp/en/research/report/021/pdf/021-22.pdf

Tang, L., & Meng, Y. (2021). Data analytics and optimization for smart industry. *Frontiers of Engineering Management*, *8*, 157–171. https://doi.org/10.1007/s42524-020-0126-0

Tang, L., & Wang, X. (2010). A Two-Phase Heuristic for the Production Scheduling of Heavy Plates in Steel Industry. *IEEE Transactions on Control Systems Technology*, *18*(1), 104–117. https://doi.org/10.1109/TCST.2009.2014960

Tasgetiren, M. F., Pan, Q.-K., Liang, Y.-C., & Suganthan, P. N. (2007). A Discrete Differential Evolution Algorithm for the Total Earliness and Tardiness Penalties with a Common Due Date on a Single-Machine. *2007 IEEE Symposium on Computational Intelligence in Scheduling*, 271–278. https://doi.org/10.1109/SCIS.2007.367701

Terence, B. (2020). *The Major Applications of Steel*. Retrieved May 23, 2023, from https://www.thoughtco.com/steel-applications-2340171

Twidale, S., Abnett, K., & Chestney, N. (2021). *EU carbon hits 100 euros taking cost of polluting to record high*. Retrieved May 23, 2023, from https://www.reuters.com/markets/carbon/europes-carbon-price-hits-record-high-100-euros-2023-02-21/

United-Enterprises. (n.d.). *Continuous Annealing Line (CAL)*. Retrieved May 21, 2023, from https://www.unient.co.in/all-products/Turnkey-Projects/continuous-annealing-line

Wegel, S., Ivanov, A., Lenz, R., & Volling, T. (2024). Scheduling of parallel continuous annealing lines with alternative processing modes to optimize efficiency under tardiness constraints. *European Journal of Operational Research*. https://doi.org/10.1016/j.ejor.2023.12.032

Williamson. (n.d.). *Annealing Line Application Overview*. Retrieved May 23, 2023, from https://www.williamsonir.com/wp-content/uploads/2018/06/Annealing_Line_Application_Note.pdf

Zhang, B., & Yang, Y. (2014). Solving production scheduling with chain constraints in parallel production lines by column generation. *Proceeding of the 11th World Congress on Intelligent Control and Automation*, 800–803. https://doi.org/10.1109/WCICA.2014.7052818

Zhao, S., & Yang, Y. (2016). Differential evolution algorithm for solving coil scheduling problem in parallel continuous annealing lines. *2016 Chinese Control and Decision Conference (CCDC)*, 3417–3422. https://doi.org/10.1109/CCDC.2016.7531573

Junior Management Science 10(3) (2025) 810-830



Junior Management Science

www.jums.academy
ISSN: 2942-1861



KPIs for Sustainability: Defining the Strategy for a Sustainable Future in the Insurance Industry

Lea Wedel

Technical University of Munich

Abstract

Sustainability transformation has gained traction across industries worldwide. Given their critical risk exposure, this transformation is of extreme precarity to the insurance industry. Yet, literature lacks a comprehensive approach for the development of a sustainability strategy covering the environmental, social and governance (ESG) dimensions. This work bridges the gap by evaluating methods for defining sustainability Key Performance Indicators (KPIs) and their integration into strategy. A roadmap is then developed that guides successful strategy implementation. Based on a systematic literature review of 5.000+ academic papers, this work features a quantitative and qualitative analysis of literature and evaluates four core papers on the definition of sustainability KPIs for the insurance industry. Leadership commitment and ESG integration into core business emerge as the most important factors towards a successful transformation. The resulting sustainability roadmap provides a blueprint for insurers to embed ESG values, enabling businesses of all sizes to participate. This work contributes to academia and industry by supporting the development of comprehensive and successfully integrated sustainability strategies.

Keywords: insurance; KPIs; strategy; sustainability; transformation

1. Introduction

In 1987, the United Nations defined sustainability as "[...] meet[ing] the needs and aspirations of the present without compromising the ability to meet those of the future" (World Commission on Environment and Development, 1987, p. 34). This corresponds to social and environmental well-being for all people. In a business context, sustainability has often been transferred into Corporate Social Responsibility (CSR) and is"[...] understood to be the way firms integrate social, environmental and economic concerns into their values, culture, decision making, strategy and operations in a transparent and accountable manner [...]" (Hohnen, 2007, p. 4).

The first time that the private financial sector became part of this sustainability movement was in 1992 and the insurance sector followed in 1997 (cf. Schanz, 2022, p. 9). Out of these endeavors, the United Nations Environment Programme Finance Initiative (UNEP-FI) emerged. The initiative catalyzes a network of more than 500 banks and insurers and works with them towards a sustainable future (cf. UNEP-FI, n.d.). To this date, UNEP-FI is still highly relevant and on the forefront of working on the global sustainability transformation. One example for that is the recently founded Net-Zero Insurance Alliance (NZIA) that commits members to reduce emissions in their underwriting portfolios to net-zero by 2050 (cf. UNEP-FI, 2023b, p. 5).

In 2015, the United Nations (UN) presented the 17 Sustainable Development Goals (SDGs) that have set an agenda of 169 targets towards achieving sustainability by 2030. They are an "[...] urgent call for action by all countries [...]" (United Nations Department of Economic and Social Affairs, 2024) to combine efforts and transform this world into a more sustainable one. The 17 goals have defined an

I want to express my gratitude to everyone who supported me during the process of writing this thesis. Special thanks go to the TUM Chair of Management Accounting for the opportunity to write my thesis with them. My sincere appreciation is further extended to Prof. Dr. Maximilian Blaschke and Max Nadicksbernd for supervising my thesis and supporting me every step along the way.

DOI: https://doi.org/10.5282/jums/v10i3pp810-830

© The Author(s) 2025. Published by Junior Management Science.
This is an Open Access article distributed under the terms of the CC-BY-4.0 (Attribution 4.0 International). Open Access funding provided by ZBW.

action plan that can be adapted to a breadth of industries and businesses. It is common practice for many insurers to align their endeavors towards a sustainability transformation along SDGs that are relevant for their business (cf. Sherwood and Sullivan (2021, p. 7) and Allianz (2021, p. 10)).

The topic of sustainable development is of extreme precarity to the insurance industry. In mere two decades we have seen the ESG movement grow "[...] from a corporate social responsibility initiative launched by the UN into a global phenomenon that is reshaping the asset management and broader business management" (Schanz, 2022, p. 9). To support this statement, PwC has shown that 85% of insurers believe that sustainability and ESG practices will influence their business along all functions (cf. 2022, p. 2).

The coming years will not see a change regarding this critical situation. Hansen, Sato and Ruedy of Columbia University have shown that climate change can be expected to accelerate much faster than previously thought (cf. 2024, p. 3). This increase in climate change has frequently been linked to an increase in severe weather events of all kinds (cf. IPCC (2023, p. 46) and European Environment Agency (2017, p. 19)) which in turn leads to an increase in insurance claims (cf. Sato and Seki, 2010, p. 333). In consequence, insurance premiums will rise and there will be a strong need for new mathematical models to predict the frequency of such events – which has recently been a huge challenge in the industry (cf. European Insurance and Occupational Pensions Authority, 2022, pp. 11–12). The last resort of insurers is to stop their operations in areas that are considered high-risk, which has been starting to happen in Florida and California, where a high number of insurers have stopped selling new policies (cf. Bogage, 2023). The same pattern will unfold in further regions if climate change is allowed to continue at its current rate.

However, sustainability is relevant to the insurance sector beyond the discussion of climate change. The social dimension is of extreme relevance as absorbing risks is the core business of insurers which in turn provides stability and security to communities and societies (cf. Schanz, 2022, p. 11). Insurers face difficult financial dynamics as a result of unstable societal and economic conditions. There is an interdependence between the insurance industry providing stability and security on the one hand and the insurance industry needing these factors to continue their operations on the other hand (cf. Trichet, 2005, pp. 67-68).

In short, climate change and related sustainability topics pose a large risk to insurers. This risk can only be mitigated by prevention measures and efforts towards the transformation into a sustainable business. As one of the most affected industries, the insurance sector should be on the forefront of these endeavors.

This paper aims to provide a comprehensive approach towards the definition of a new sustainability strategy in the insurance industry. It aspires to bridge a gap between extensive literature on ESG risks on one hand and their implementation and strategy integration in the insurance industry on the other hand. The consideration of all three ESG dimensions allows for the identification of interconnections and a view on their combined impact on strategy while existing literature often considers these dimensions in isolation. A detailed evaluation of the three ESG dimensions will be carried out and translate theoretical sustainability concepts into specific and strategic KPIs. In accordance with literature, investment management will not be covered given its dominance in existing works (cf. Stricker et al., 2022, p. 3). The paper concludes with the presentation of a step-by-step roadmap that enables the strategic integration of sustainability measures into the core of insurance companies and their strategic orientation.

This work underlines an urgent need for a new strategy that will guide firms through a new era of sustainability risks, requirements and opportunities. Following the concept of a systematic literature review, more than 5.500 papers were evaluated quantitatively and qualitatively. Based on this research, a comprehensive set of KPIs and a roadmap that will serve as a blueprint to companies was developed. The methodology will be explained further in chapter 2.

After concluding the description of the process that led to the identification of the four core papers analyzed further throughout this body of work, a quantitative analysis of the literature review will be provided. The strategic relevance and exponentially growing importance of ESG related topics will be highlighted based on purely quantitative data. This will be followed by a short introduction of the four key papers that stresses their importance for this work and outlines the current state of literature.

In chapter 4, three core papers on the evaluation and implementation of KPIs into company strategy will be evaluated, compared and synthesized with regard to the insurance industry. The relevance of each dimension will be substantiated with a quantitative analysis of the dataset. Next, an evaluation of a fourth core paper concerning the successful introduction of these KPIs in companies to ensure long-term strategic success concerning ESG topics will be provided.

Chapter 5 will present a distinctive roadmap that is meant to enable insurance companies to include sustainability measures into their daily business and their strategic orientation by following a step-by-step guide.

Finally, a future outlook will be provided and further fields of research will be identified.

2. Method of data collection and evaluation

This work is based on a systematic literature review. The methodology was chosen as it "offers a broader and more accurate level of understanding than a traditional literature review" (Pati & Lorusso, 2018, p. 15). The goal of this research was the identification of sustainability related trends in the insurance industry through covering an extensive time horizon and to provide a future outlook based on available literature. The maximization of the number of input papers was the preferred option to fulfill this task.

In line with the approach of Tranfield, Denyer and Palminder (cf. 2003, pp. 214-219), the process of this system-

atic review was broken up into three main steps. First, a thorough planning of the process was carried out, followed by the actual conduction of the review and completed by an organization of the results (cf. Tranfield et al., 2003, pp. 214-219). The main step of the planning process was the definition of two keyword lists that provided the basis for the review. In line with the focus of this thesis on the connection between sustainability strategy and the insurance industry, two lists of keywords in the respective dimensions of sustainability and insurance were drafted.

The keyword definition process happened in iterative steps. After designing a first draft of 20 keywords in both lists, their usefulness was evaluated by scanning a number of Whitepapers. This step ensured that relevant and current language was used to cover the breadth of available information. After further consultation with industry experts from a reputable German insurance company, two keyword lists of 21 and 31 keywords, respectively, were finalized.

Lastly, the next steps were defined. A python algorithm was used to iteratively combine each keyword of the insurance dimension[1] with each keyword of the sustainability dimension[2]. Finally, the identified papers were rated based on the most used keywords over all abstracts and then sighted manually in a last step.

With the finalization of a step-by-step plan for the review, the planning step was completed. Next, the review had to be carried out systematically.

Orgeolet et al. (cf. 2019, pp. 815-817) have proven the usage of artificial intelligence (AI) to support the manual implementation of a systematic literature review to be very efficient. The increasing number of published papers and consequently the increasing number of relevant papers makes it difficult to obtain a broad picture of the current state of literature without the support of AI technologies. For this reason and in compliance with current regulations, the conducted research was supported by the use of AI, namely the programs "ChatGPT"[3] and "Connected Papers"[4].

The first step towards the execution of the review was drafting a python script that would enable the systematic extraction of data from the web. With this script, the iterative combination of keywords from the sustainability and the insurance list was carried out in a Google Scholar[5] search.

For each combination, the algorithm opened the first 15 detail pages of the Google Scholar search results and extracted predetermined information into an excel file. In the excel file, the information was structured in columns by title, author, journal, abstract and the year of publishing. To ensure sufficient data quality, each abstract was copied into the file manually.

Including the built-in function of the algorithm to prevent duplications, the final total of papers in the excel file amounted to 6.396 papers. The duplication measures of the algorithm proved to be insufficient, so a further removal of duplicates by abstract, title and URL was carried out and supported by ChatGPT. This led to a total of 5.505 papers in a cleaned version of the excel file. Several quantitative analyses carried out on this data that will be evaluated in chapter 3.

To quantitatively and qualitatively analyze the papers, various mechanisms were used. First, several different methods of tagging and subsequently rating the abstracts were implemented. ChatGPT supported this tagging process by scanning all abstracts for the provided tag-words and entering them in a new column. This process was repeated multiple times and with different tag-words to provide coverage of different focus areas. For each method of tagging, two rating approaches were used. First, the rows were sorted by number of individual tags and second, they were ordered by the total count of tags including repeated occurrences in the abstracts. This ensured that the rating process was not biased by the length of the abstracts of the papers.

Finally, the first 20 papers of all tagging approaches were sighted repeatedly to decide on the keywords that would yield the best results. Consequently, the following 14 keywords were defined for the final method of tagging: Sustainability, Insurance, Risk, Social, Environmental, Governance, Financial, Climate, Investment, Strategy, Data, Innovation, Compliance and KPIs.

After the exclusion of all papers with zero tags, the first 50 papers of both rating methods were then sighted manually and evaluated into categories ranging from "not useful to this work" to "highly accurate core papers". The final category of core papers included four papers, while a high number of supporting papers contributed to this work. A quality check of the core papers using different ratings was purposefully delayed until this step to ensure that the total amount of publications was used for the quantitative analysis.

Figure 1 illustrates the methodology applied to collect, filter and analyze the data for this work.

As a second-to-last step, the four remaining core papers were quality checked by consulting three tools. Two different VHB-Ratings were used, namely the "VHB-Rating 2024 für Publikationsmedien"[6], the VHB-JOURQUAL 3[7] and additionally the ABDC Journal Quality List[8]. One of the four core papers was not listed in either one of these tools but a plausibility check led to the decision to include this paper. It was published by the Geneva Association, an internationally recognized association of insurance and reinsurance CEOs (cf. The Geneva Association, n.d.). Two of the remaining three

[1] Keyword List of the Insurance Dimension: [Investments, Insurance, Retail, Industry, Motor, Home, Flood, Catastrophe, Wildfire, Governance, Risk, Risk Management, Life, Health, Own Operations, Reporting, Finance, Strategy, Security, Ehrbarer Kaufmann, Responsibility]

[2] Keyword List of the Sustainability Dimension: [Sustainability, Net-Zero, Engagement, Exclusion, Climate, Biodiversity, Science-Based Targets, Diversity, Equity, Inclusion, Transition, Mitigation, Impact, Governance, Compliance, Anti-Money Laundering, Human Rights, Ethics, SDG, Social, Diversity, ESG, Green, Impact, Carbon, Environmental Management, CSR, Greenwashing, Waste, Local]

[3] chatgpt.com

[4] connectedpapers.com

[5] scholar.google.com

[6] https://vhbonline.org/service/vhb-rating-2024

[7] https://vhbonline.org/service/vhb-jourqual/vhb-jourqual-3

[8] https://abdc.edu.au/abdc-journal-quality-list/

Keyword List Insurance
Investments, Insurance, Retail, Industry, Motor, Home, Flood, Catastrophe, (wild)fire, Governance, Risk, Risk Management, Life, Health, Own Operations, Reporting, Finance, Strategy, Security, Ehrbarer Kaufmann, Responsibility



Keyword List Sustainability
Sustainability, Net-Zero, Engagement, Exclusion, Climate, Biodiversity, Science-Based Targets, Diversity, Equity, Inclusion, Transition, Mitigation, Impact, Governance, Compliance, Anti-Money Laundering, Human Rights, Ethics, SDG, Social, Diversity, ESG, Green, Impact, Carbon, Environmental Management, CSR, Greenwashing, Waste, Local



Input	Database	Output	Number of articles
Iterative combination of keywords from both lists	Google Scholar	Excel file with one row per paper	6.396



Unique documents	Elimination of duplicates by title, abstract and URL	Number of articles	5.505



Further exclusion of documents leading to final sample size	Elimination of papers with 0 tags	Number of articles	4.379



Number of final core papers	4	Number of identified supporting papers	9

Figure 1: Implementation of search process using Python and Google Scholar

papers yielded ratings of A and B from the ABDC Journal Quality List. One core paper was only included in the 2015 Rating of VHB where it received a rating of C, which still corresponds to a recognized and professional journal, hence it was included in this work.

Not all supporting papers were rated with these tools but a high number of recurring journals was observed which allowed for a quick plausibility check.

Finally, to guarantee the inclusion of all relevant existing literature, the four identified core papers were analyzed with the AI tool Connected Papers. This tool takes a so-called *origin paper* as an input and then analyzes 50.000 publications to build a graph that is based on co-citations and the coupling of bibliographies (cf. Behera et al., 2023, p. 3). While no further core paper was found with this approach, a number of supporting papers were identified by analyzing the graphs.

This step concluded the process of data collection. A qualitative analysis of the complete dataset of all papers followed by an introduction of the core papers will be carried out in chapter 3. Chapter 4 will build up on this with a comparative analysis of the core papers and some supporting papers while chapter 5 will provide a combination of both to identify trends and a roadmap.

3. Quantitative and qualitative analysis of the literature

In the following section a quantitative analysis of the dataset will be provided. The chapter will be concluded by an introduction of the identified core papers. The relevance of each core paper will be stressed in relation to the research question.

3.1. Quantitative findings

To ensure a holistic picture, the dataset was not just analyzed qualitatively but also quantitatively. Several statistical analyses were conducted to visualize development over time and identify current and future trends. These trends will be presented in the following chapter. The results from the quantitative data analysis support the qualitative findings presented in the introduction and underline the need for action by the insurance industry.

First, the frequency of published academic papers over all papers was analyzed which is displayed in Figure 2. With the first paper on social insurance dating back to 1913, a drastic increase over the last 100 years is observed. A takeoff happens around the year 2000 and since then the number of published papers has been growing exponentially. Similar trends can be observed for the number of published sustainability reports by companies that are following the Global

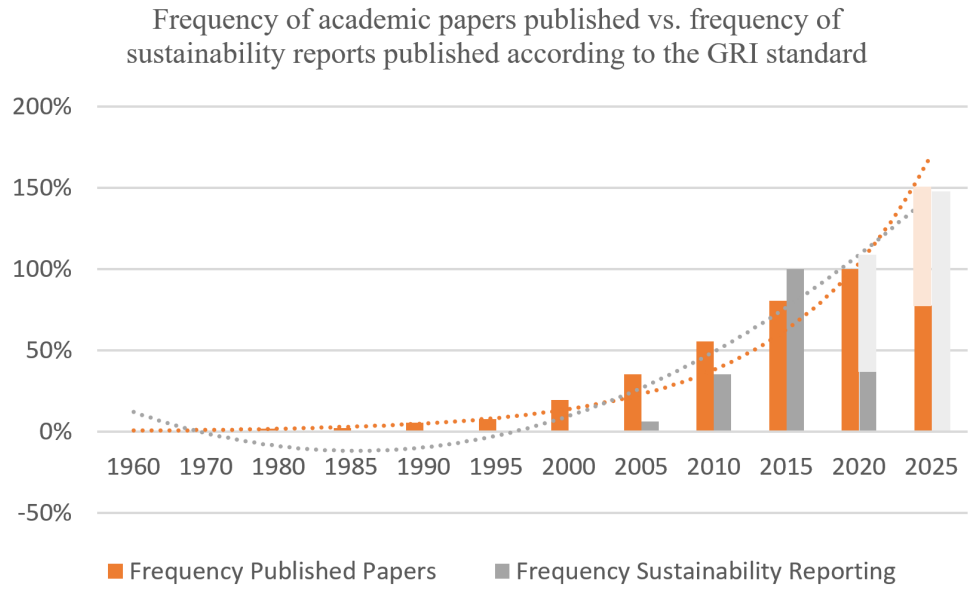


Figure 2: Frequency distribution of published academic papers vs. released sustainability reports

Reporting Initiative (GRI)[9] standards based on data by Verdantix (cf. 2017). A drastic increase can be seen from 2010 onwards and continues to develop in a polynomial curve. The progress regarding published sustainability reports happens in a time-delay compared to the frequency of total published papers. This underlines that we are still in the early days of sustainability adoption in the industry. Acting now will provide companies with a head start and allow them to leverage an early mover advantage in the following years.

A similar picture is painted when looking at the keywords "strategy" and "sustainability" and the frequency by which they jointly occur in academic papers as seen in Figure 3. The first time these words appeared together was in 1999 and in the following years only occasional mentions were recorded. A more dynamic development started in 2015 and has not halted since. There is no reason to assume a decrease in the upcoming years which supports the prediction of an expected takeoff in sustainability endeavors in the academic and industrial world.

In this regard, a special focus should be shown upon the changes in the conversation on sustainability and sustainable development over the course of the last 15 years. Part of this can be seen in the wording that is used in academic literature. Figure 4 shows the difference in frequency of the 10 most frequent keywords found in all abstracts. The graph shows their absolute frequency in occurrence in the years 2010-2019 compared to their absolute frequency in the years 2020-2024. Notably, the second time period is less than half the years of the first, which underlines the drastic message of this graph. Increases can be seen for almost all keywords, with climate having an increase of more than 100 additional mentions in all abstracts. There are only three decreases, "ethical", "biodiversity" and "governance". Biodiversity and governance are almost at a change of 0 so this can presumably be attributed to the difference in time spans. The decrease of the keyword "ethical" on the other hand is most likely of other nature. While in the last decade, sustainability was more of an ethical topic, a question of morale, we are now in a time of it being the main narrative. The commitment to sustainability is not optional anymore, the only question is the level of commitment.

Lastly, Figure 5 shows a word cloud that is the result of a frequency analysis of all papers regarding their abstracts using Python. The prominence of the words "impact" and "risk" shows the strong awareness for the fact that climate change and sustainability topics have the potential to disrupt and disturb entire industries and pose an immense risk if not mitigated correctly.

Further, the words "model", "system" and "framework" show that the need for a new strategy that guides firms through these challenges is very present. Sustainability risks need to be addressed and targeted systematically along the whole value chain and cannot be limited to certain areas or industries.

In combination, these analyses strongly suggest the need for literature and research on valid strategies and systems to navigate the complex landscape of sustainable development for insurers. The qualitative analysis of the literature will build up on this data in the next chapters.

3.2. General characteristics of the core papers

In the following section a short introduction of the key papers will be provided. First results and a classification as well as the rhetoric importance of the papers will be explained.

[9] www.globalreporting.org/about-gri/



Figure 3: Frequency of the occurrence of the keywords "strategy" and "sustainability" in the same publication

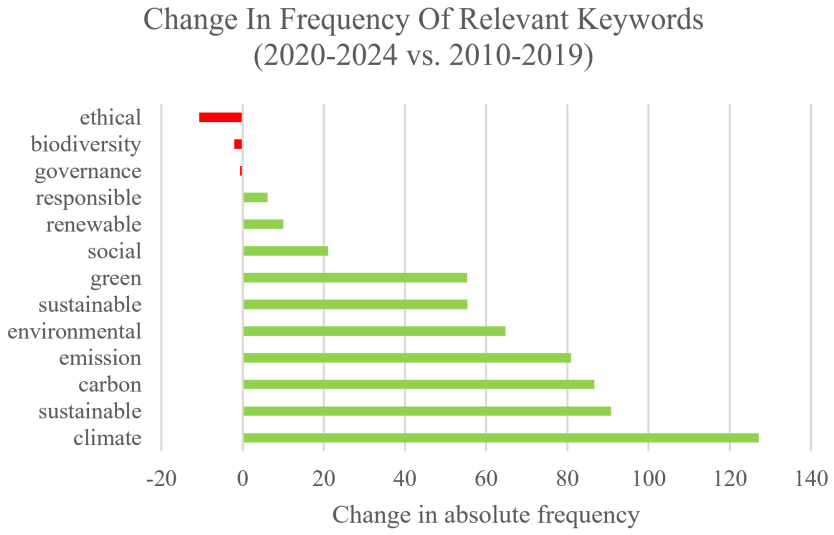
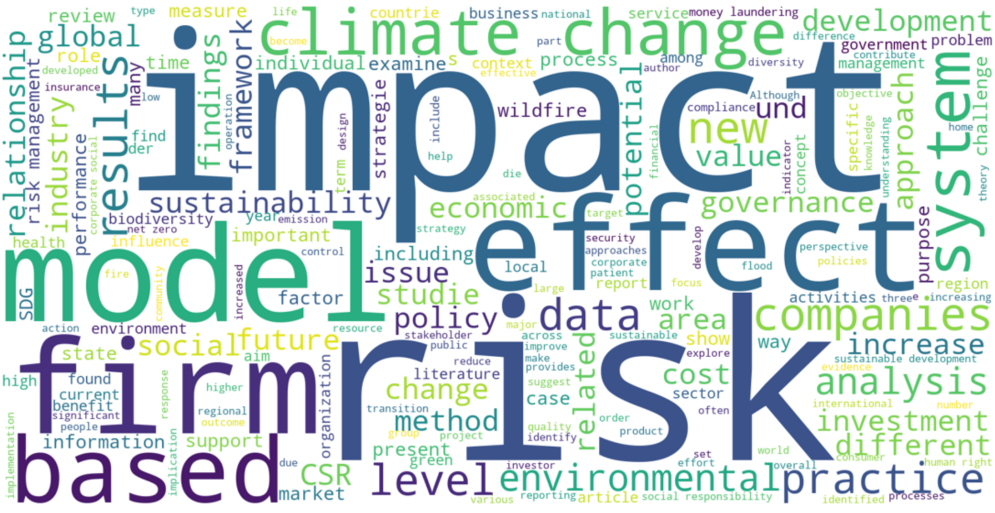


Figure 4: Change in frequency of relevant keywords from 2010-2019 to 2020-2024


impact
climate change
model
effect
risk
firm
based
system
sustainability
economic
governance
development
companies
analysis
investment
different
practice
environmental
level
data
policy
studie
factor
issue
results
findings
framework
relationship
global
review
measure
business
service
money laundering
government
problem
management
approach
challenge
potential
value
new
wildfire
including
important
influence
biodiversity
net zero
emission
year
health
SDG
industry
performance
risk management
social
future
state
environment
aim
high
found
current
benefit
significant
information
support
CSR
market
organization
method
change
literature
present
green
related
cost
sector
case
increase
way
show
area
work
activities
purpose
region
increasing
three
perspective
policies
flood
focus
report
local
security
approaches
corporate
patient
target
strategie
term
design
include
help
die
associated
strategy
control
fire
increased
community
resource
stakeholder
public
across
improve
make
provides
suggest
reduce
transition
group
project
quality
identify
order
product
often
effort
set
world
consumer
overall
investor
due
various
reporting
article
social responsibility
identified
processes
human right
sustainable development
evidence
international
number
explore
sustainable
major
action
implementation
implication
response
outcome
regional
higher
operation
low
many
find
der
insurance
developed
role
time
individual
examine
countrie
process
context
effective
author
among
diversity
difference
part
become
life
type
compliance
Although
objective
Specific
Knowledge
understanding
home
theory
indicator
concept
financial
develop
target
security
approaches
corporate
patient
large
work
activities
three
increasing
region
purpose
perspective
policies
flood
focus
report
local
including
issue
studie
factor
policy
data
area
companies
increase
way
show
cost
sector
case
related
analysis
investment
different
practice
environmental
level
CSR
market
organization
method
change
literature
present
green
sustainable

3.3. Core paper on KPIs defined by a two-fold approach

The first paper that was identified as a core paper is an article by Hristov and Chirico (cf. 2019) evaluating potential KPIs used in sustainability strategies by using a two-fold approach. First, they conducted a systematic literature review with the goal of identifying KPIs that enable the development of a comprehensive sustainability strategy (cf. Hristov and Chirico, 2019, p. 2). Second, they drafted a questionnaire and interviewed 25 Italian managers with more than 5 years of experience in the area of sustainability. These managers were asked about their strategic goals and the prioritization of the previously identified KPIs. This two-sided approach allowed them to set up a sustainability perspective containing strategic goals, KPIs and measurements for the environmental, social and economic dimension which can be found in Appendix 2 (cf. Hristov and Chirico, 2019, p. 2). Their use of the economic dimension can reasonably be translated into the governance dimension of the ESG rationale. It represents the strategic handling of the integration of KPIs of the environmental and social dimension as well as cost handling and enhancement of the sustainability transformation process. For the purpose of this work, the KPIs identified in the economic dimension will therefore be evaluated in the governance section.

The resulting sustainability perspective of the study is based on a concept called the sustainability balanced scorecard that aims to integrate "[...] environmental and social issues [...] with the general management of a business unit" (Figge et al., 2002, p. 281). This scorecard will allow companies to have a readily available blueprint which they can extend and fill with specific goals and measurements according to the needs of their companies (cf. Hristov et al., 2019, p. 13). This makes the model of the sustainability balanced scorecard applicable beyond the boundaries of specific industries and hence of high relevance for the insurance sector and this work.

3.4. Core paper on a roadmap for executive management

The second core paper identified discusses a roadmap towards green insurance for executive management (cf. Stricker et al., 2022). The roadmap is laid out along the classical insurance value which can be seen in Figure 6 (see Porter (1985) as cited in Eling and Lehmann (cf. 2018, p. 362), and Rahlfs (cf. 2007, pp. 169-170)). It divides the activities of insurance companies into primary activities such as product development and claim management on one side and support activities such as general management and Information Technology (IT) on the other side.

The paper focuses on three elements along the value chain of insurers, which are first product development as well as marketing and sales, second risk management and underwriting and third operations and claims management (cf. Stricker et al., 2022, p. 3). It reviews current literature with regard to these three dimensions and suggests a set of metrics for monitoring the sustainability efforts of companies.

The paper further identifies ESG risks that can be transferred into KPIs to promote sustainability. These risks as displayed in Table 1 provide a structured approach for general management to evaluate exposure to their firm and derive strategic goals and KPIs based on their identified risks. While the paper has a strong focus on environmental KPIs along the value chain, social and governance aspects will also be considered shortly.

The final roadmap developed in this paper can be found in Appendix 3 to ensure completeness.

This paper is of high relevance to the research objective of this work as it pinpoints clear KPIs along the value chain of insurance companies which provides a different approach than the first core paper. They complement each other in achieving a synthesis of a comprehensives set of KPIs with a focus on environmental activities.

3.5. Core paper on the social impact of insurers

The third core paper by Schanz (cf. 2022) concentrates on the social dimension of sustainability and briefly covers some governance aspects. It concludes the triad of core papers on specific KPIs and ensures that all three ESG dimensions have sufficient coverage. The paper refers to social sustainability as the "capacity of current and future generations to live and work in healthy and liveable conditions that promote diversity and equal opportunities" (Schanz, 2022, p. 6). It further calls out the relevance of the insurance industry by pointing to the socially beneficial core of their business (cf. Schanz, 2022, p. 6).

The report was issued by the Geneva Association and is based on more than 40 interviews with executives and experts (cf. Schanz, 2022, p. 8). It follows an explanatory approach examinating insurers in the social dimension of sustainability and aims to provide a framework for guiding firms towards social sustainability. The findings of the report have been summarized into four main recommendations. First, a three-tier approach for the management of social sustainability that consists of the maximization of social impact, protection of benefits and exploration of further actions should be implemented. Second, the current business model has to be reviewed followed by third, the adaption of core business activities. Lastly, governance towards the social dimension has to be integrated into top management activities (cf. Schanz, 2022, p. 8).

The paper takes the SDGs related to social sustainability as a basis and provides suggestions for the introduction of KPIs and activities stemming from these. Lastly, the paper provides some suggestions for the introduction of governance KPIs into the strategy that will help to promote social sustainability on a long-term basis. While there is no focus on the environmental dimensions, the governance aspects mentioned are very generalist which makes them applicable for both, the social and environmental dimension. This makes the paper highly relevant for the current research question as it explores areas that were not covered by the other works and allows a synthesis towards a comprehensive sustainability transformation strategy.

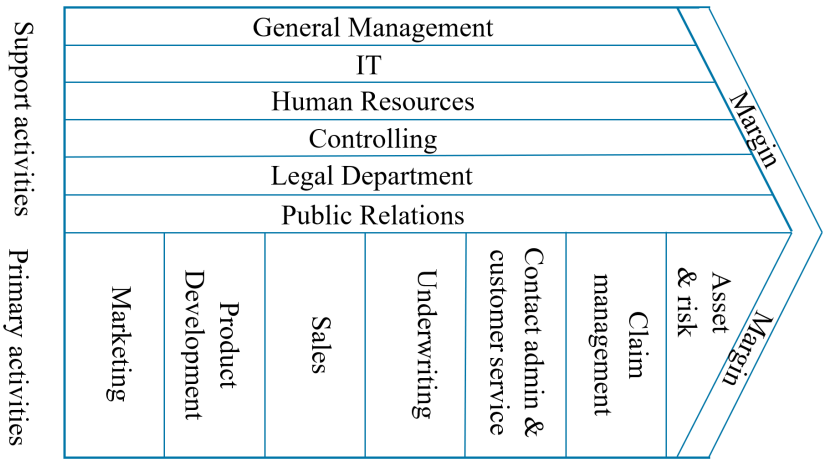


Figure 6: Classical value chain of insurance companies based on Porter (1985) as cited in Eling and Lehmann (cf. 2018, p. 362) and Rahlfs (cf. 2007, pp. 169-170)

Issues	Description
Environmental	• Climate mitigation and adjustment to climate change • Protection of biodiversity • Sustainable use and protection of water and maritime resources, sustainable land use • Transition to a circular economy, avoidance of waste and recycling • Avoidance and reduction of environmental pollution, protection of healthy ecosystems
Social	• Compliance with recognized labor standards (no child labor, forced labor or discrimination) • Compliance with employment safety and health protection • Appropriate remuneration, fair working conditions, diversity, training/development opportunities • Trade union rights and freedom of assembly • Guarantee of adequate product safety, including health protection • Application of the same requirements to entities in the supply chain • Inclusive projects and consideration of the interests of communities and social minorities
Governance	• Tax honesty and anti-corruption measures • Sustainability management by the board, board remuneration based on sustainability criteria • Facilitation of whistle-blowing, employee rights guarantees • Data protection guarantees and information disclosure

Table 1: Characterization of ESG Risks based on BaFin (cf. 2020, pp. 12-15), COSO and WBCSD (cf. 2018, pp. 9-10), EBA (cf. 2020, pp. 30-31) and EIOPA (cf. 2019, pp. 27-29)

3.6. Core paper on systematic integration into strategy

The last core paper chosen for this analysis focuses less on specific KPIs and more on a framework for the successful integration of said KPIs into a sustainability strategy (cf. Johannsdottir and McInerney, 2018). The study aims to "[...] bridge the gap between strategy, strategy implementation and sustainability by identifying the critical factors evident in strategies and activities [...]" (Johannsdottir & McInerney, 2018, p. 1253).

The research approach was divided into several steps. First, a case study with the largest non-life insurance companies in the Nordic Region was conducted, followed by a study on medium- sized companies in Iceland (cf. Johannsdottir and McInerney, 2018, p. 1257). The last step was a verification of the proposed framework by literature review. These steps concluded in the development of a so-called 5 C Framework (cf. Johannsdottir and McInerney, 2018, p. 1257). The five C's represent the five steps towards successful strategy implementation. They are "1) commitment, 2) configuration, 3) core business, 4) communication and 5) continuous improvements" (cf. Johannsdottir and McInerney, 2018, p. 1258).

The paper will be used as a final step in this work to demonstrate the successful integration of sustainability measures into sustainability strategy. Taken together with the evaluation of the results of all ESG dimensions, it will provide the basis for the subsequent development of the roadmap towards successful strategy implementation in chapter 5.

3.7. Further lines of research and placement of work

Further lines of research exist mostly in the area of the growing influence of ESG-related topics on the insurance industry in the years to come. The discussion of specific KPIs is rare and usually concludes with a statement stressing the need for further research.

Additionally, there is a high number of studies evaluating the risk that stems from the disregard of sustainability factors. The environmental dimension is discussed the most in this stream of literature, as the increase in natural disasters poses an enormous risk for the insurance industry.

The social dimension with regard to the definition of specific KPIs and frameworks is the one least discussed in research. Most academic papers on the social dimension deal with the role of insurance in developing countries and their important contribution to social welfare. The specific discussion of KPIs or transformation strategies is rare, if at all found.

This work contributes to existing literature in summarizing and synthesizing efforts made to define and evaluate different KPIs. It provides a combination of sets of KPIs for all three ESG dimensions together with a framework to implement these KPIs into a sustainability strategy. This ensures that efforts are not halted at the definition but a transformation of the business is enabled for insurance companies.

4. Findings

This chapter will identify and evaluate different KPIs and strategies mentioned in the core papers. They will be categorized based on the ESG dimensions. First, the environmental factors will be investigated as the most researched area in literature. Second and third, social and governance factors will be discussed. This will be concluded by an evaluation of strategy integration of all three dimensions which intends to provide an overarching picture. The results of this chapter have been used to develop a holistic framework presented in chapter 5. The framework is expected to serve not only as a starting point for firms to identify potential KPIs for a sustainability strategy but also allow the personalization and tracking of identified goals.

4.1. Environmental factors

The environmental dimension of the ESG narrative is generally seen as the easiest to grasp and quantify. Improving the environmental performance comes down to reducing greenhouse gas (GHG) emissions, which ultimately means reducing carbon dioxide (CO_2) emissions (cf. Lee et al., 2015). Compared to the other ESG dimensions of social and governance these emissions are quantifiable in almost all operations.

CO_2 and GHG emissions are commonly divided into three categories, so-called scopes. Scope 1 comprises all direct emissions that are owned and controlled by a company, scope 2 emissions are indirect emissions from electricity consumed by a company and lastly scope 3 emissions are all remaining indirect emissions that occur as a consequence of the company's activities but cannot be directly controlled by them (cf. World Business Council for Sustainable Development and World Resources Institute, 2004, p. 25). With increasing reporting requirements, firms will have to report their emissions up and down their value chain which is intended to exert pressure on them and their value chain (cf. Wagenhofer, 2024, p. 6). These requirements will force companies to increasingly disclose their emissions which constitutes a need for them to develop emission reduction strategies to stay competitive in the market and fulfill customer demand for sustainable practices (cf. Wagenhofer, 2024, p.3).

Figure 7 paints a similar picture and confirms the need for a strategy handling emissions. Although the frequency of the occurrence of the combination of the keywords "strategy" and "environment" seems to have come to a halt in recent years, it needs to be noted that the last bin covers only 3.5 years instead of 5 years. This is due to the analysis being conducted in mid-2024. That means that we are and still will be seeing an increase in the occurrence of this combination. The development of emission handling strategies that will adequately cover the environmental dimension is an ongoing process. One reason for a less prominent increase than with other keywords might be that the debate on the environmental dimension has been going on for a longer time. Yet, there is still much potential for improvement and acting now will benefit the firms in the future.

4.1.1. Clustering by strategic goals

According to Hristov and Chirico (cf. 2019, p. 7), the most relevant and prevalent KPIs in the environmental dimension are related to GHG emissions, renewable resources, resource consumption and waste. Table 2 shows the goals and related indicators that have been identified as being the most important measures in their interviews (cf. Hristov and Chirico, 2019, p. 9). The topic of renewable energy has clearly been identified as the most prevalent issue with 100% of managers ranking it as the most important one. All managers further agreed that the renewable sources rate should be used to measure the progress towards this goal. The further goals identified as top priority in declining frequency were the reduction of superfluous materials, the reduction of GHG emissions, safeguard governmental policies and lifestyles towards natural resources. In Appendix 1, a compilation of subsequently identified specific KPIs related to these goals can be found. The list contains several KPIs per goal which demonstrates that the identification and agreement upon goals is the most difficult part, whereas the definition of KPIs to measure the progress is a quicker process.

An important aspect to note is how much specific KPIs will differ among companies. One example for this is the goal of reduction of superfluous materials seen in Table 2. The associated KPI is the rate of efficiency of materials used. The identification and agreement upon the clear reduction goal is the difficult part, the KPI stems from it naturally. Every company will have different materials use, so the exact KPI for reduction will differ notably while the overarching goal

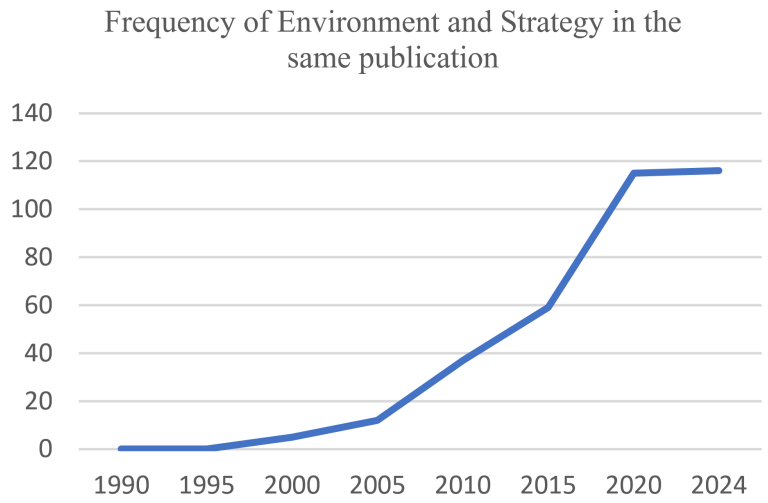


Figure 7: Frequency of the occurrence of the keywords "environment" and "strategy" in the same publication

Goal	Frequency	%	KPI	Frequency	%
Renewable Energy Sources	25	100	Renewable Sources Rate	25	100
Reduction Superfluous Materials	23	92	Efficiency Materials Use Rate	23	92
Reduction Greenhouse Gases	22	88	Emissions Greenhouse Gases Rate	18	72
Environment's Safeguard Governmental Policies	16	64	Waste Reduction Rate	16	64
Lifestyles towards natural resources	16	64	Recycled Materials used rate	16	64

Table 2: Frequency distribution of KPIs in the environmental dimension (cf. Hristov and Chirico, 2019, p. 10)

will be of use for a broad spectrum of companies in different industries.

4.1.2. Clustering along the value chain and based on risks

Stricker et al. (cf. 2022, p. 4) take a different approach into the definition of KPIs by addressing them along the value chain. In this, they acknowledge that the insurance business as a whole has to be part of the sustainability transformation and undertake a more structured approach. In contrast to Hristov and Chirico (cf. 2019) they focus on the environmental dimension and only shortly cover social and governance aspects.

In their evaluations of the environmental aspect, they focus on emission reduction by transforming different operational activities into their related emissions and potential for action. They clearly apply the GHG scope model to this narrative. In this, they summarize all identified goals under the umbrella term of CO_2 emissions.

For scope 1, Stricker et al. (cf. 2022, p. 12) consider onsite heating emissions and fleet emissions, scope 2 consists of the share of renewable energy and scope 3 comprises emissions from travel, emissions from suppliers and emissions from employee commutes. While this approach might be harder to grasp at first, the clustering of emissions into three scopes is going to be increasingly important due to new reporting requirements and allows for clear communication.

The main difference between the goals identified by Hristov and Chirico (cf. 2019, p. 10) versus Stricker et al. (cf. 2022, p. 12) is the decision to structure them. While the development of a model for clustering into scopes requires a higher effort, a clear framework will also ease the compliance with reporting requirements in the future. This makes the approach by Stricker et al. (cf. 2022, p. 12) preferable.

Additionally, Stricker et al. (cf. 2022, p. 6) also cover green insurance products which extends the range of KPIs. While "insurance is a promise and as such an intangible product" (Stricker et al., 2022, p. 5) that does not produce emissions in itself, the sustainability of a product is linked to the object that is insured and the operations of the company. According to Stricker et al. (cf. 2022, p. 5), a sustainability roadmap for insurance products can be defined by four major aspects: insurance for green objects, inclusion of green policy features, sustainable transition claims support and pro-

motion of environmentally friendly behavior. All of this is enabled through a regular dialogue with customers. Examples for green policy features in cases where environmental impact cannot be reduced is CO_2 compensation (Stricker et al., 2022, p. 6). An example for this is a CO_2 compensation extension for travel insurance. The potential for green insurance products is huge for companies, especially as there is no clear definition for these products which allows extensive tailoring by insurers. A market leader position can be inherited by allocating time and resources to the development of these products. Potential ideas range from pay-as-you-drive offers for car insurance policies to re-use incentives as claim handling strategies to only name two examples (cf. Stricker et al., 2022, p. 6).

In further literature it has been established (cf. Marti et al., 2024, p. 207) that "product development evolution linked to ESG criteria is going further in the practitioner world than in the academic one". This statement has been confirmed in the process of the literature review. There is a gap in research with regard to sustainable insurance products. The market would benefit greatly from a detailed roadmap targeting the development and consecutive establishment of green insurance products.

To return to the bigger picture it is important to note that there is no "one fits all" approach to a greener insurance. A comprehensive approach by insurance companies is needed to retain credibility and prevent greenwashing accusations (cf. Stricker et al., 2022, p. 6). This entails a sustainable orientation along the whole value chain ranging from own operations to risk management and claims handling.

4.1.3. Discussion on environmental factors

Both papers provide valuable insights into the environmental dimension and appropriate KPIs to quantify progress towards a sustainable insurance business. Specific policies covered in the papers have not been discussed as single examples administered across different organizations will not yield a comprehensive insight. A quantitative analysis or empirical study over a broad range of insurance companies would bridge an important gap in this area.

It is important to note, that the overarching KPI of reducing CO_2 emissions –by the achievement of net zero or an emission-cutting process – is omnipresent in literature. Much debate concerns the exact measures needed to achieve this goal and the part that compensation has to play. A general trend is a distancing from compensation and efforts towards actual emission cutting as the avoidance of greenwashing accusations becomes increasingly important for firms (cf. Lyon and Montgomery, 2015, pp. 223-224).

The most important aspect that can be derived from this analysis is the strategy for the development of KPIs. The definition of strategic goals should be the overarching directive based on which KPIs can be derived. This provides a structured approach instead of the random assignment of KPIs related to specific climate issues. Additionally, the development of green insurance products will be critical over the next years. Giving customers the option to make a sustainable choice while upholding a high-quality service is the baseline. The identification of core areas for the implementation of these products will allow companies to foresee pain points and stay competitive in the future.

4.2. Social factors

The social dimension has been discussed frequently in literature, yet there is no clear definition for the scope of social impact. Most definitions cover a broad range of aspects and consider the social dimension as a capacity for providing welfare and life-enhancing living conditions to citizens (cf. Hristov and Chirico (2019, p. 7) and McKenzie (2024, p. 12)). The review of the papers obtained by the literature review concludes that the social dimension is the least researched dimension of all three when it comes to the implementation and evaluation of specific KPIs. This is most likely also linked to limited understanding of the social dimension as "[...] social factors are less tangible and come with limited data on how they can impact a company's performance" (Schanz, 2022, p. 10).

While this might lead to the conclusion that activities with regard to environmental factors are more important to companies, an analysis of the keywords "social" and "strategy" paints a different picture. Similar trends as shown above for the environmental dimension can be seen in the combination of the keywords "social" and "strategy" in Figure 8. A slow development seems to start around the year 2000 and up until now there is no sign of a significant slowdown[10]. This underlines the importance of considering the social dimension in the movement towards sustainability. According to some reports, the developments regarding the social dimension have been gaining even more traction since the start of the global Covid-19 pandemic and the ongoing conflict in Ukraine as the social imbalances caused by this will force businesses to pay more attention to social matters (cf. Schanz, 2022, p. 15).

The increasing importance of the social dimension will additionally be driven by increasing regulations and legislative initiatives worldwide. Some examples of recent years include the European Union (EU) taxonomy regulation, the Corporate Sustainability Reporting Directive (CSRD) and a potential social EU taxonomy (cf. Schanz, 2022, pp. 15-16).

An interesting perspective for the approach of the social dimension has been provided by the Geneva Association. They developed a model that applies the 3 scope model of the environmental dimension to the social dimension of ESG as can be seen in Figure 9 (cf. Schanz, 2022, p. 27). This approach partitions the broad field of social impact that a firm has into three scopes. Scope 1 represents the social impact that is directly controlled by the firm, scope 2 is the impact that an insurer has through its own operations and employees and scope 3 is the impact that results from the insurance

[10] This is again based on the consideration of different time brackets as this analysis was concluded in mid-2024.

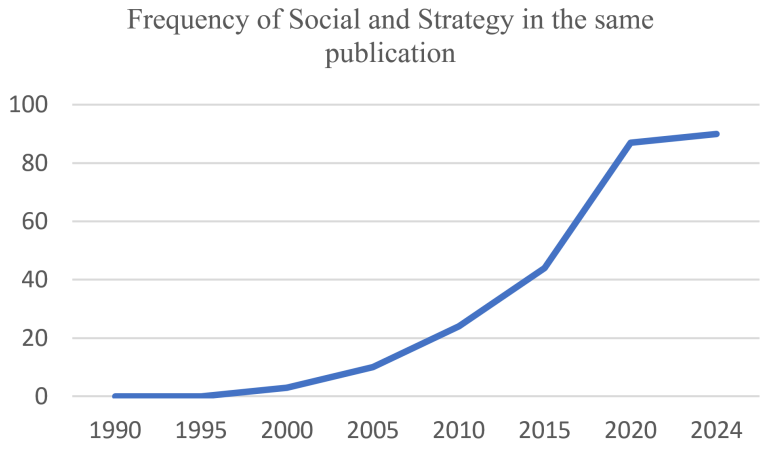


Figure 8: Frequency of the occurrence of the keywords "social" and "strategy" in the same publication

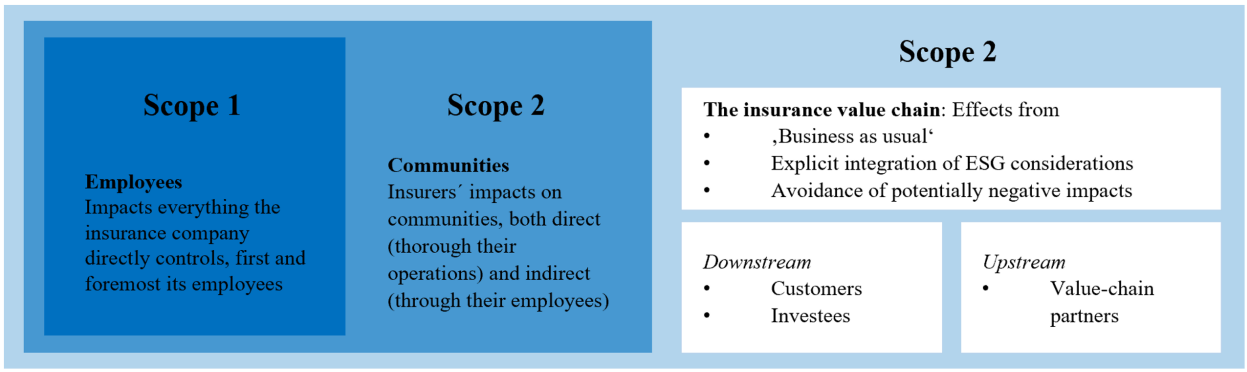


Figure 9: The 3-Scope Model applied to the social dimension (cf. Schanz, 2022, p. 27)

value chain. As in the environmental model, scope 3 cannot be directly controlled although accounting for the biggest part of social impact (cf. Schanz, 2022, p. 27). This approach allows for a clustering of social impact that will serve to structure KPIs in the social dimension.

4.2.1. Clustering by strategic goals

Similar to the structuring of the environmental KPIs, Hristov et al. (cf. 2019, pp. 8-9) have structured their discussion on social indicators along the three steps of goal identification, KPI identification, and lastly KPI ranking through interviews. The results of this process are displayed in Table 3.

The strategic goals upon which Hristov and Chirico (cf. 2019, p. 8) agreed were the encouragement of employees to accept cultural change, the improvement of working conditions, the guarantee of respect for human rights and the participation in social initiatives. Again, from these strategic goals specific KPIs were derived and ranked by priority in interviews. The first observation is the reduced agreement on specific goals and KPIs compared to the environmental dimension. This probably stems from the fact that the social dimension is broader and less clearly defined. This makes it harder to measure progress and therefore set specific goals. The most unity can be seen in the goal “Inclusion of all society members in the development of sustainable goals” with 21/25 managers agreeing on this goal. The associated KPI is the integration rate of employees. This KPI is directly linked to the goal of “Inclusion of employees in the decision process” that was prioritized by 15 out of the 21 respective managers. While both KPIs are hard to survey, they will provide very valuable information to the company and serve as a quantitative KPI that can be consequently reported.

When considering the results of the research by Hristov et al. (cf. 2019, p.10) under the approach of the scope model, a focus on scope 1, i.e. the employees, is observed. There is some focus on scope 2 but close to no consideration of the value chain impact of scope 3. To date, scope 3 impact can rarely be found in literature, but this can be expected to change soon. With the Act on Corporate Due Diligence in Supply Chains[11] in Germany (cf. Federal Ministry of Labour and Social Affairs, n.d.) and similar acts in the European Union (cf. Directorate-General for Communication European Commission, 2024) companies will be required to increasingly disclose information about social and environmental factors along their value chain.

[11] Also known as Lieferkettensorgfaltspflichtengesetz (LkSG)

Goal	Frequency	%	KPI	Frequency	%
Inclusion of all society members in the development of sustainable goals	21	84	Integration rate	21	84
Stakeholder's relationship	18	72	Stakeholder's satisfaction rate	15	60
Participation to the social initiative	18	72	Total participation social initiative national and local level	14	56
Inclusion of employees in the decision process	15	60	Employee integration rate	11	44
Social marketing policies	14	56	Customer satisfaction index	10	40

Table 3: Frequency distribution of KPIs in the social dimension (cf. Hristov and Chirico, 2019, p. 10)

4.2.2. Clustering by SDG-related fields

With the goal of covering all three scopes of social impact a different approach was taken by Schanz (cf. 2022, pp. 22-23). He clustered the areas of social impact into actions related to the dimensions of people and prosperity. Figure 10 shows the SDGs that were identified in relation to the people dimension and have laid out the areas of strategic goals. The SDGs were divided into three sub-fields of the people dimension, namely dignity and equality, health and well-being and an overarching people pillar. Table 4 shows the core metrics that are proposed to measure social impact in all three categories. An example is the ratio of Chief Executive Officer (CEO) compensation to average employee compensation in the dignity and equality bucket. The KPIs proposed serve as an inspiration that can be adapted by insurers to fit their specific needs and business model. They are not ranked or presented in a specific order but rather represent an incomprehensive list of suggestions.

The second dimension that was further analyzed by Schanz (cf. 2022, p. 23) with regards to social impact is prosperity which is related to the SDGs 1, 8, 9 and 10 as displayed in Figure 11. This dimension is characterized by the creation of long-term value by companies towards the benefit of employees and society (cf. Schanz, 2022, p. 23). There are fewer specific metrics attached to this dimension but suggestions include the turnover rate of employees, diversity indicators and the economic value a company generates and distributes. For insurers, this is specifically related to the payment of claims and benefits to customers and the development of better products and services (cf. Schanz, 2022, p. 23).

The last aspect mentioned with regard to social prosperity is the investment in communities and society (cf. Schanz, 2022, p. 23). There are many opportunities to do this that range from special tax payments to the funding and sponsoring of social or employee initiatives. This is also consistent with the goal of social marketing policies by Hristov et al. (cf. 2019, p. 10). This endeavour is of specific importance as it serves two purposes. One, it generates social impact and two, it allows a very strategic positioning of the company in the market towards the customer and the own employees. There are few other goals that have the same amout of leverage and traction while holding a strong balance between reputation and impact.

4.2.3. Discussion on social factors

The clustering along the impact areas of SDGs instead of strategic goals allows for a structured and holistic approach. In combination with the scope model for social impact, a distinct and clear process can be defined that covers the areas of social impact in depth and clarity. Ideally, companies should take the SDGs related to social sustainability as a basis and then meticulously work out strategic goals for all three scopes of social sustainability. The presentation of strategic goals by Hristov et al. (cf. 2019, p. 10) will help to navigate a balance between economic prosperity and social impact. In combination, the synthesis of these two papers serves as a blueprint for the development of social sustainability KPIs by insurance companies.

4.3. Governance factors

Governance factors are generally thought of as being the hardest to grasp and understand. Several metaphors exist to describe the relevance of governance factors, all with loss of accuracy. The dimension of governance is more all-encompassing and can be seen as a structuring and more strategic element incorporating both, environmental and social aspects. Figure 12 shows the development of the occurrence of the combination of the keywords "governance" and "strategy" in academic literature and strongly underlines the trend towards the integration of governance processes in strategy. The steeper increase of the graph quantitively underlines the need for a comprehensive approach to the governance dimension by insurance companies.



Figure 10: SDGs in relation to the social dimension of sustainability (cf. World Economic Forum, 2020)

Dignity and Equality	• The percentage of employees per employee category by indicators of ethnicity • The ratio of the basic salary and remuneration for each employee category by significant locations of operation for women to men, minor to major ethnic groups, and other relevant equality areas • Ratios of standard entry-level wage by gender compared to local minimum wage • Ratio of the annual total compensation of the CEO to the median of the annual total compensation of all its employees • An explanation of the operations and suppliers considered to have significant risk for incidents of child labor, forced or compulsory labor
Health and well-being	• The number and rate of fatalities as a result of work-related injury • High-consequence work-related injuries (excluding fatalities) • An explanation how the organization facilitates worker's access to non-occupational medical and healthcare services • The scope of access provided for employees and workers
People pillar	• Average hours of training per person undertaken by the company's employees during the reporting period, by gender and employee category • Average training and development expenditure per full time employee

Table 4: Proposed core metrics related to the identified SDG dimensions based on Schanz (2022, p. 22)



Figure 11: SDGs related to social prosperity (cf. World Economic Forum, 2020, p. 37)

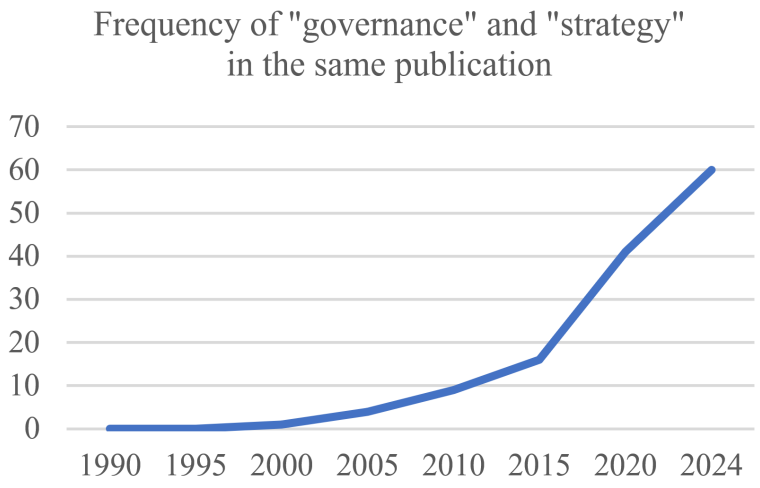


Figure 12: Frequency of the occurrence of the keywords “governance” and “strategy” in the same publication

Hristov et al. (cf. 2019, p. 9) use the same approach as before to identify clear strategic goals for the governance[12] dimension. They were not able to define specific strategic goals and instead described the strategic mission as the “ [...] capacity to create durable growth of the main economic indicators [...] [while] sustain[ing] the population, enhancing territorial uniqueness and efficiently employing resources [...]” (Hristov & Chirico, 2019, p. 9).

Table 5 shows the specific goals and KPIs derived from this underlying mission. Notably, more agreement than in the social goals and KPIs can be observed. All managers agreed on the prioritization of the goal “Trend of the financial indicators” with the linked KPI “ROI related to sustainable investments”. This is apparent as this goal is designed to measure and improve economic output and growth. While sustainable behavior is increasingly important to companies, they also need to make sure that their underlying fundamentals are strong to ensure stable operations in the future.

The remaining goals and associated KPIs are mostly related to quality aspects with links to product, marketing, training and communication. The mention of revenues invested in training provides an important approach that is prevalent in academic literature. It is one of the few ways to quantitatively prove to the market that you are not only applying measures towards social and environmental impact in your operations, product development etc. but also shifting the mentality of your firm and educate your personnel.

Another relevant aspect is the updating of policies towards ESG measures mentioned by Stricker et al. (cf. 2022, p. 14). They stress the importance of policies that are reviewed and updated to reflect sustainability issues. This is especially important for insurers considering their business model. Overall, insurers deal with risks in all kinds of categories. That means that the inclusion of potentially disrupting ESG risks across all policies will serve to increase the stability of the business and ensure long-term success.

While these KPIs are covering specific measures that could be undertaken by insurers, Schanz (cf. 2022, p. 38) has taken a different approach and evaluated the governance dimension as an enabler for KPIs from other dimensions[13]. He stresses the importance of embedding the orientation towards sustainability at the top-level management, the board of directors and senior executives. Further, he recommends the appointment of ESG market leads and the enablement of responsible decision-making (cf. Schanz, 2022, p. 38). While not clearly stated, this entails the allowance of room for difficult decisions which might require investments that will only pay out on a long-term basis and more importantly, it allows for the commitment of leaders to sustainability topics.

Supporting literature has shown that some insurers go even further and commit to the appointment of a Chief Sustainability Officer (CSO) who streamlines and coordinates ESG efforts (cf. Sherwood and Sullivan, 2021, pp. 3-10). And while this is a clear commitment, research has also shown that these CSOs tend to struggle with similar hurdles as the ones that have been described before. Observations include a lack of clear definitions, metrics and resources (cf. Sherwood and Sullivan, 2021, pp. 3-10). The overcoming of these aspects will be discussed in the next section on the successful integration of sustainability measures into strategy and further summarized in the proposed roadmap in section 5.

A last, highly debated and critical, aspect that Schanz (cf. 2022, p. 38) mentions is the linkage of top management compensation with targets related to sustainability. This is confirmed by Stricker et al. (cf. 2022, p. 16) who state that

[12] Hristov et al. (2019) have worked with the three dimensions environmental, social and economic. In chapter 3.3 an explanation is provided for the placement of the economic dimension int the governance dimension.

[13] Schanz (2022) focuses on the social dimension of sustainability. As he regards the governance dimension as an enabler, all measures will be used in this paper generically for both, the environmental and the social dimension.

Goal	Frequency	%	KPI	Frequency	%
Trend of the financial indicators	25	100	ROI related to sustainable investments	25	100
Products quality	24	96	Check on products quality	21	84
Revenues invested in training, R&D, Marketing and Communication	20	80	Revenues invested rate	15	60
Communication technology process	20	80	Ratio of investments to it, infrastructures and web on total sales	14	56
The quality of the process	18	72	Sustainability and quality of the post-sale technical assistance rate	12	48

Table 5: Frequency distribution of KPIs in the governance dimension (cf. Hristov and Chirico, 2019, p. 10)

the integration of sustainability into the leadership culture of the company is key. Board remuneration based on sustainability criteria is identified as one of the levers to achieve this (cf. Stricker et al., 2022, p. 8). Linking compensation to sustainability incentives has been a highly discussed topic for years and will likely gain more traction in the future. While not necessarily common practice at this time, this can be expected to change rather soon. Creating a strategy in regard to this allows companies to gain a competitive advantage and establish credibility among their stakeholders and in the market.

4.3.1. ESG in risk management

One of the main aspects of sustainability governance is the risk management. Stricker et al. (cf. 2022, pp. 9-10) have developed a roadmap for the risk management of insurance companies. This roadmap has been transferred into a graphical representation shown in Figure 13. Each bullet point mentioned can be transferred into a KPI for general management. As an example, the hiring of staff with expertise in sustainability risks (bucket 1) can be linked to a number of new hires or a percentage of total staff having this expertise. Similarly, the inclusion of sustainability risks in internal and external sustainability reporting (bucket 4) is to be translated into a number of risks and factors that are accounted for.

The formulation of these risks is consistent with further streams of literature and is one of the most researched areas with regard to sustainability. Schanz (cf. 2022, p. 38) is confirming the importance of the inclusion of ESG dimensions into risk strategies in his paper at the governance level as one of the main enablers of successful sustainability practices.

While the roadmap in Figure 13 does not specify KPIs, it leads a way for companies to include these steps into their core business. The exact formulation of KPIs will be up to the respective company but in that, the framework underlines its usefulness. It can be freely adapted to the needs of each individual insurer while making sure that a holistic approach is taken and the risk management is sufficiently covered in a sustainability strategy.

4.3.2. ESG in underwriting

Lastly, the underwriting practices of insurance companies should be evaluated with a focus on the governance dimension. The decision to include this into the generic part of governance was taken as underwriting allows no clear clustering into the environmental or social dimension but should rather take into account all three ESG dimensions.

Similar to their roadmap for risk management (Figure 13), Stricker et al. (cf. 2022, p. 10) have also developed a roadmap for underwriting practices of insurance companies. They divide the process into five key steps. First, a thorough ESG due diligence has to take place with experts from different focus areas that identify ESG risks and reputational issues. Second, specific criteria and metrics for the assessment of ESG risks have to be defined. It is suggested to consult external sources like UNEP-FI (cf. 2023a) or Allianz (cf. 2021). Third, the inclusion of sustainability in the assessment of risks is proposed (cf. Stricker et al., 2022, p. 10). That refers to the inclusion of ESG factors but also the provision of tools to access information concerning the information related to ESG risks that are relevant to underwriting processes. Next, a clear decision-making process should be defined. That also includes thresholds for decision making in single cases and escalation procedures. And lastly, ESG expertise should be delivered to customers. This is targeted towards the dialogue but also insurance solutions that are supporting the sustainable transition of the company (cf. Stricker et al., 2022, p.10).

The sustainable transformation of underwriting practices is largely focused on environmental issues given that the cost to the insurance sector that is caused by severe weather events could exceed 1 trillion United States Dollar (USD) per year by 2040 (cf. Dlugolecki, 2008, p. 87). Yet, it has also been stressed that the consideration of social factors is highly

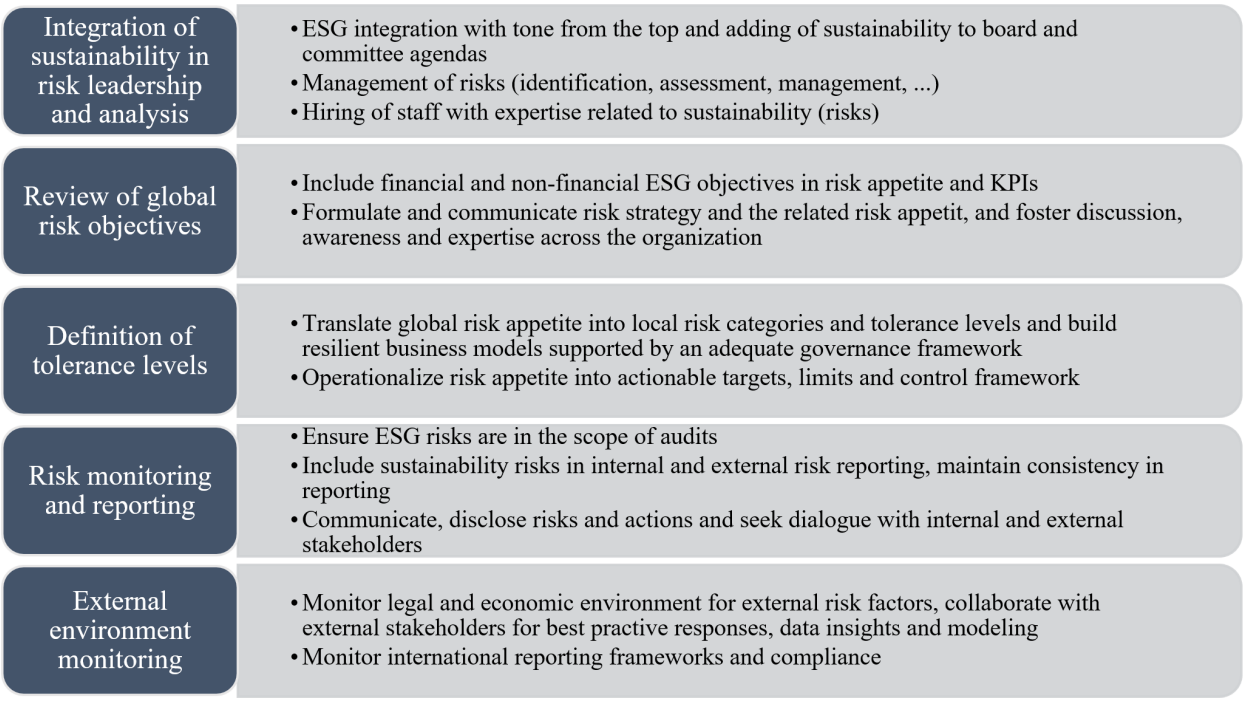


Figure 13: Roadmap for risk management based on Stricker et al. (cf. 2022, pp. 9-10)

relevant. Schanz (cf. 2022, p. 7) calls for action on the algorithms that underwriting practices are based on to ensure that the risk of exclusion of certain customer groups is managed. The paper even establishes challenges related to the social impact of underwriting as one of the main challenges in the daily business of insurers in the next five to ten years (cf. Schanz, 2022, p. 25).

Underwriting practices regarding a transformation towards sustainability are still limited today. Yet, it has been repeatedly established that the integration of ESG factors is increasingly taking place (cf. Marti et al., 2024, p. 204). Taking action now will allow a competitive edge in the market and bears potential for a market-leading position.

4.3.3. Discussion on governance factors

The governance dimension of ESG is an incredibly broad one. Governance is seen as a strategic element that allows the positioning of a company in the market and enables the integration of environmental and social aspects into the business of insurers. The strategic KPIs and goals should therefore always be aligned with business fundamentals.

One important measure to highlight is the training and qualification of employees as well as the creation of sustainability-related roles as a way to credibly signal the importance of sustainability topics within the operations of your firm to the market and competitors. The integration of sustainability throughout the whole company ties into this argument. It entails the integration into policies but also the involvement of top management and the possible linkage of management or board remuneration with ESG performance.

Lastly, the fields of risk management and underwriting have been examined in greater detail. A clear roadmap and specific KPIs in these areas will ensure the resilience of the business.

In conclusion, comprehensive governance mechanisms that are transparent and cover all aspects of the company value chain, strategic goals and risk categories will enable a competitive edge. Implementing these policies with the necessary finesse and credibility will be a challenge for years to come but will also allow companies to take a market-leading position if they start focusing on the execution now.

4.4. Integration of ESG dimensions into strategy

As a last step, insurers need to integrate the identified KPIs into their "[...] culture, core business, strategy and structure" (Johannsdottir & McInerney, 2018, p. 1253). It is commonly understood that "executing strategy is just as important, if not more important, than formulating strategy" (Higgins, 2005, p. 3). This stresses the importance of companies having access to a comprehensive roadmap for integration after the successful development of a strategy.

Barriers that firms see themselves confronted with range from structural dimensions over political, human and cultural dimensions to dimensions that are out of a company's control such as regulation and market conditions (cf. Stewart et al., 2016, pp. 25-26.

To ensure that companies are successful in the integration of sustainability practices by strategy, Johannsdottir and McInerney (cf. 2018) have developed the so-called 5 C Framework to implement sustainability strategies in insurance companies.

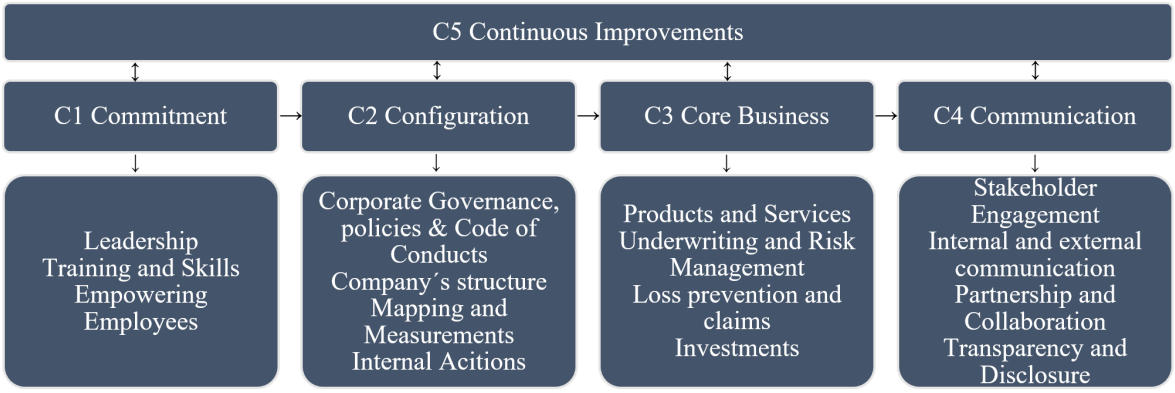


Figure 14: Insurer's Five C Framework of implementing environmental sustainability strategies (cf. Johannsdottir and McInerney, 2018, p. 1259)

The general process can be seen in Figure 14. The first step towards successful integration of sustainability is the definite commitment of the leaders. It is their "[...] role to establish environmental strategy, vision, mission and goals" (Johannsdottir & McInerney, 2018, p. 1259). The second step is the configuration of the business. That refers to the integration of ESG factors into corporate governance and the company structure as well as the steering of internal actions concerning sustainability (cf. Johannsdottir and McInerney, 2018, p. 1260). The third aspect suggests that to maximize their impact, companies will have to integrate sustainability into their core business (cf. Johannsdottir and McInerney, 2018, p. 1260). The last driver that is integrated vertically is internal and external communication. This can be through a special spokesperson, partnerships or concise sustainability reports (cf. Johannsdottir and McInerney, 2018, pp. 1260-1261).

The last dimension of this framework is C5 – Continuous Improvements. It integrated horizontally and stresses the need for constant monitoring and reevaluation of sustainability measures and risks (cf. Johannsdottir and McInerney, 2018, p. 1261). The integration into business structures is of utmost importance as sustainability is a highly complex field that requires periodic reassessments.

This framework is a valuable roadmap for all insurance companies trying to integrate sustainability into their day-to-day business and core values. Defining measures and steps for all five dimensions will enable the successful integration on a long-term basis.

5. Discussion and development of a strategy roadmap

To conclude this work, a roadmap towards a sustainable strategy in the insurance industry has been created. The thorough research of all three dimensions of ESG within this work has been used to develop a comprehensive roadmap which can be seen in Table 6. It provides a clear framework within which companies will be able to steer towards a sustainable future. This is aimed to bridge a gap between the development of strategic goals, the definition of KPIs and the subsequent integration into strategy.

The roadmap has been divided into six strategic steps, namely the preparation and commitment, the research and benchmarking, the definition of KPIs, the definition of measurement and targets, the implementation and integration into strategy and lastly, the monitoring and continuous improvement. The last step is to be interpreted as a continuous loop that can trigger the reiteration of specific strategic steps or the start of a new process. The roadmap further defines the actions, the responsible stakeholders and the desired outcome for the goals of each strategic step. It therein provides a fully holistic approach that can be adjusted and personalized by each company. The definition of the triad of action, responsibility and outcome further ensures that companies can monitor their progress before entering the next step. If necessary, they are able to revisit specific steps instead of rushing through the process prematurely.

While the roadmap does not feature specific strategic goals or KPIs, it aims to provide a practical implementation guide for insurance companies seeking to transform their business. The definition of KPIs is and will remain highly customized to each organization. A more streamlined set of goals and indicators can be expected to develop in the coming years, but until then, this roadmap serves as an enabler for insurance companies of all sizes to take their own steps towards a more sustainable future.

6. Conclusion and outlook

The relevance of the topic of sustainability to the insurance industry has been stressed repeatedly throughout all chapters. Managing risks is fundamental to the business of insurers and being able to predict these risks forms the core. Naturally, this makes them one of the most exposed industries with regard to sustainability risks and reinforces the role they should be taking on as market leaders in the ongoing sustainability transformation.

Strategic Step 1: Preparation and Commitment			
	Action	**Responsibility**	**Outcome**
ESG vision and strategic goals	Define and outline ESG vision and strategic goals of the company	Top management (and key stakeholders in the company)	ESG vision and mission statement that will serve as a guidance
Leadership Commitment	Top-level management and board have to plead commitment to ESG initiatives	CEO and board of directors	Formal statement confirming commitment of time and resources
Appointment of ESG leaders	Allocation of roles, ESG team appointment	Top management	Clearly defined ESG team leading the process
Strategic Step 2: Research and Benchmarking			
	Action	**Responsibility**	**Outcome**
Benchmarking and Review	Review of industry standards, best practices and peer performance	ESG (research) team	Extensive understanding of current ESG practices, benchmarking with peers and identification of gaps & room for improvement
Strategic Step 3: KPI Definition			
	Action	**Responsibility**	**Outcome**
Identification of strategic goals for each ESG dimension	Definition of strategic goals for environmental, social and governance dimension	ESG team in cooperation with appointed leaders from respective departments	Strategic goals in the environmental, social and governance dimension that sets a framework for the consecutive definition of specific KPIs
Development of clearly defined KPIs for each strategic goal	Definition of specific KPIs based on the strategic goals and the mission statement	ESG team in cooperation with departments and cross-functional teams	KPIs for the environmental, social and governance dimension
Strategic Step 4: Measurements and Targets Definition			
	Action	**Responsibility**	**Outcome**
Definition of Measurement	Establish clear methods to quantitatively and qualitatively measure the KPIs	ESG team and respective departments	Transparent process of measuring the KPIs (can be published in sustainability report if desired)
Setting of targets	Based on the KPIs and the process of measuring them, define targets and timeline for them	ESG team in cooperation with respective departments and top management in next step	Tables with environmental, social and governance targets (see Appendix 2)
Strategic Step 5: Implementation and Integration			
	Action	**Responsibility**	**Outcome**
Integration of KPIs into Daily Business Processes	Integrate the KPIs into daily operations and business planning, link them to performance management	All departments under the guidance of the ESG team	Transformation into a sustainable business
Development of Reporting and Communication Framework	Creation of a framework for reporting and ESG communication with internal and external stakeholders	ESG team with reporting and communication departments	Clearly defined reporting framework and communication channels & timeline
Strategic Step 6: Monitoring and Continuous Improvement			
	Action	**Responsibility**	**Outcome**
Monitor performance to review and adjust KPIs	Regularly scheduled monitoring of market + periodic performance review to adjust KPIs and targets.	ESG team with relevant departments. Adjustment of KPIs in cooperation with top management	Continuous improvement and periodic performance reports
Engagement of Stakeholders	Engagement with stakeholders to collect feedback on ESG policies and ensure involvement of employees	ESG team and if required specific departments	Enhanced relationships with internal and external stakeholders

Table 6: Roadmap towards the sustainable transformation of insurers (own illustration)

To be able to achieve this positioning, guidance and research is needed so that insurance companies of all sizes are able to succeed in the adaption of ESG measures. But while there has been a high amount of research on the topic of sustainability there is still a large gap regarding concrete steps towards realizing a sustainable transformation. This work contributes to bridging this gap by providing a framework that companies will be able to freely adapt and tailor while ensuring that a holistic picture will be kept at all times.

Through a systematic literature review encompassing more than 5.500 papers, KPIs for all three ESG dimensions have been evaluated quantitatively and qualitatively. Combined with a consideration of strategy implementation, an action plan has been developed that helps firms to identify strategic goals, related KPIs and their measurements. As a last step, the framework enables the successful integration of measures and KPIs into strategy.

Regarding the environmental dimension, it has been found that the reduction of CO_2 emissions is the main strategic goal from which all kinds of measures and KPIs will stem. For this dimension, the consideration of credibility and the avoidance of greenwashing policies is of high importance as mere compensation of emissions is not sufficient anymore.

The social dimension is less straightforward and harder to grasp. It has been established that evaluating the social impact based on the related SDGs allows for a structured approach. Combined with the adapted scope model (see Figure 9) a clear guideline for the definition of strategic goals and KPIs is provided. Exploring a similar approach of using the related SDGs as a basis for the definition of environmental KPIs like Allianz has been doing (cf. 2021, p. 10) would be of great value for future research.

Lastly, the governance dimension is potentially the most critical. Without proper integration of sustainability values into policies across the organization and embedding of sustainability values throughout the organization, efforts towards a sustainable transformation are likely to fail. Board and top management remuneration is an important aspect that will gain traction in the future and provides credibility to internal and external stakeholders.

Positioning the firm in the market as a sustainability leader is not an easy task but given the current traction, this will be a differentiating factor in the coming years. Sustainability has always been a long-term concern which makes it harder for firms to commit to it but in the light of recent developments this can be expected to change. Additionally, increasing reporting requirements and customer demands will further force companies into a transformation and already having appropriate systems in place will be of huge advantage by then.

Future research should be conducted in the area of implementation and evaluation of KPIs. Long-term and empirical studies tracking the ESG performance of companies over several years are needed. The precise measurement of acceptance, usefulness and impact on economic performance of applied KPIs is critical. Currently, no comprehensive evaluation of these factors exists, especially in regard to insurance companies. Conducting studies on this topic will be highly valuable to industry peers and prove useful within and beyond the insurance sector.

Building up on this, a set of follow-up case studies on the application of the framework that has been developed within this work is needed to evaluate its usefulness in the market. The definition of KPIs will always need to be customized but it is crucial to investigate the usefulness of these framework in guiding a sustainability transformation.

While the literature that has been reviewed had no clear geographic focus, it mostly covers examples and research of countries considered industrial nations. This limits the applicability of the developed frameworks and identified KPIs to developing countries. Further research on the unique situation that insurers face in developing countries is critical in this regard. The precarity of the social and environmental dimensions should be closely monitored in the next years. It is essential that comprehensive guidelines on managing sustainability risks in developing countries are established in the coming years.

Despite the inclusion of an extensive amount of academic literature in this review, the rapid evolution of sustainability topics may have led to the disregard of relevant studies. Future research using similar approaches will therefore be appreciated in this highly dynamic field. Tailoring the keyword lists to specific research objectives will be a good starting point for evaluating the different dimensions more closely.

While this work provides a starting point towards a sustainability transformation in the insurance industry, it cannot be considered as a single source of truth. The conduction of future research will help to identify focus areas and allow the further development of the roadmap.

References

Allianz. (2021). Allianz Group Sustainability Report 2021: Building Confidence in tomorrow. Retrieved May 27, 2024, from https://www.allianz.com/content/dam/onemarketing/azcom/Allianz_com/sustainability/documents/Allianz_Group_Sustainability_Report_2021-web.pdf

BaFin. (2020). Merkblatt zum Umgang mit Nachhaltigkeitsrisiken. Retrieved June 24, 2024, from https://www.bafin.de/ref/19598538

Behera, P. K., Jain, S. J., & Kumar, A. (2023). Visual Exploration of Literature Using Connected Papers: A Practical Approach. *Issues in Science and Technology Librarianship*, *104*, 1–8.

Bogage, J. (2023). Home insurers cut natural disasters from policies as climate risks grow. Retrieved June 17, 2024, from https://www.washingtonpost.com/business/2023/09/03/natural-disaster-climate-insurance/

COSO & WBCSD. (2018). Enterprise Risk Management: Applying enterprise risk management to environmental, social and governance-related risks. Retrieved June 17, 2024, from https://docs.wbcsd.org/2018/10/COSO_WBCSD_ESGERM_Guidance.pdf

Directorate-General for Communication European Commission. (2024). Corporate sustainability due diligence. Retrieved June 19, 2024, from https://commission.europa.eu/business-economy-euro/doing-business-eu/corporate-sustainability-due-diligence_en

Dlugolecki, A. (2008). Climate Change and the Insurance Sector. *The Geneva Papers*, *33*, 71–90.

Eling, M., & Lehmann, M. (2018). The Impact of Digitalization on the Insurance Value Chain and the Insurability of Risks. *The Geneva Papers on Risk and Insurance—Issues and Practice*, *43*, 359–396.

European Banking Authority. (2020). EBA Discussion Paper on management and supervision of ESG risks for credit institutions and investment firms. Retrieved June 16, 2024, from https://eba.europa.eu/sites/default/files/document_library/Publications/Discussions/2021/Discussion%20Paper%20on%20management%20and%20supervision%20of%20ESG%20risks%20for%20credit%20institutions%20and%20investment%20firms/935496/2020-11-02%20%20ESG%20Discussion%20Paper.pdf

European Environment Agency. (2017). Climate change, impacts and vulnerability in Europe 2016. Retrieved June 22, 2024, from https://www.eea.europa.eu/publications/climate-change-impacts-and-vulnerability-2016

European Insurance and Occupational Pensions Authority. (2019). EIOPA's Technical Advice on the integration of sustainability risks and factors in the delegated acts under Solvency II and IDD. Retrieved June 15, 2024, from https://www.eiopa.europa.eu/publications/technical-advice-integration-sustainability-risks-and-factors-solvency-ii-and-insurance-distribution_en

European Insurance and Occupational Pensions Authority. (2022). European insurers' exposure to physical climate change risks. Retrieved June 12, 2024, from https://www.eiopa.europa.eu/publications/discussion-paper-physical-climate-change-risks_en

Federal Ministry of Labour and Social Affairs. (n.d.). Supply Chain Act. Retrieved May 28, 2024, from https://www.csr-in-deutschland.de/EN/Business-Human-Rights/Supply-Chain-Act/supply-chain-act.html

Figge, F., Hahn, T., Schaltegger, S., & Wagner, M. (2002). The Sustainability Balanced Scorecard - Linking Sustainablity Management to Business Strategy. *Business Strategy and the Environment*, *11*, 269–284.

Hansen, J., Sato, M., & Ruedy, R. (2024). *Global Warming Acceleration: Causes and Consequences* [Working Paper, Columbia University 2024].

Higgins, J. M. (2005). The Eight 'S's of Successful Strategy Execution. *Journal of Change Management*, *5*, 3–13.

Hohnen, P. (2007). *Corporate Social Responsibility: An Implementation Guide for Business*. IISD. Retrieved June 23, 2024, from https://www.iisd.org/system/files/publications/csr_guide.pdf

Hristov, I., & Chirico, A. (2019). The Role of Sustainability Key Performance Indicators (KPIs) in Implementing Sustainable Strategies. *Sustainability*, *20*, 1–19.

Hristov, I., Chirico, A., & Appolloni, A. (2019). Sustainability Value Creation, Survival, and Growth of the Company: A Critical Perspective in the Sustainability Balanced Scorecard (SBSC). *Sustainability*, *7*, 1–19.

IPCC. (2023). Climate Change 2023 Synthesis Report. Retrieved June 14, 2024, from https://www.ipcc.ch/report/ar6/syr/downloads/report/IPCC_AR6_SYR_LongerReport.pdf

Johannsdottir, L., & McInerney, C. (2018). Developing and using a Five C framework for implementing environmental sustainability strategies: A case study of Nordic Insurers. *Journal of Cleaner Production*, *183*, 1252–1264.

Lee, K.-H., Min, B., & Yook, K.-H. (2015). The impacts of carbon (CO2) emissions and environmental research. *Int. J. Production Economics*, *167*, 1–11.

Lyon, T. P., & Montgomery, A. W. (2015). The Means and End of Greenwash. *Organization & Environment*, (2), 223–249.

Marti, A., Bastida-Vialcanet, R., & Marimon, F. (2024). A systematic literature review: ESG criteria implementation in the insurance industry. *Intangible Capital*, *20*, 193–214.

McKenzie, S. (2024). *Social Sustainability: Towards some definitions* [Working Paper, Hawke Research Institute 2004].

Orgeolet, L., Foulquier, N., Misery, L., Redou, P., Pers, J.-O., Devauchelle-Pensec, V., & Saraux, A. (2019). Can artificial intelligence replace manual search for systematic literature? Review on cutaneous manifestations in primary Sjögren's syndrome. *Rheumatology*, (4), 811–819.

Pati, D., & Lorusso, L. N. (2018). How to Write a Systematic Review of the Literature. *Health Environments Research & Design Journal*, (1), 15–30.

PwC. (2022). Next in Insurance - ESG: A growing sense of urgency. Retrieved June 14, 2024, from https://www.pwc.com.cy/en/services/assets/pwc-next-in-insurance-esg.pdf

Rahlfs, C. (2007). *Redefinition der Wertschöpfungskette von Versicherungsunternehmen*.

Sato, M., & Seki, M. (2010). Sustainable Business, Sustainable Planet—A Japanese Insurance Perspective. *The Geneva Papers*, *35*, 325–335.

Schanz, K.-U. (2022). The Role of Insurance in Promoting Social Sustainability. *The Geneva Association*. Retrieved May 13, 2024, from https://www.genevaassociation.org/sites/default/files/2022-11/social_sustainability_report.pdf

Sherwood, D., & Sullivan, K. B. (2021). Building a more sustainable insurance industry. *Deloitte Insights*. Retrieved June 12, 2024, from https://www2.deloitte.com/jp/en/pages/financial-services/articles/ins/inclusive-and-sustainable-insurance.html

Stewart, R., Bey, N., & Boks, C. (2016). Exploration of the barriers to implementing different types of sustainability. *Procedia CIRP*, *48*, 22–27.

Stricker, L., Pugnetti, C., Wagner, J., & Röschmann, A. Z. (2022). Green Insurance: A Roadmap for Executive Management. *Journal of Risk and Financial Management*, (5), 1–19.

The Geneva Association. (n.d.). About us. Retrieved May 13, 2024, from https://www.genevaassociation.org/about-us

Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a Methodology for Developing Evidence-Informed Management Knowledge by Means of Systematic Review. *British Journal of Management*, *14*, 207–222.

Trichet, J.-C. (2005). Financial Stability and the Insurance Sector. *The Geneva Papers*, *30*, 65–71.

UNEP-FI. (n.d.). About Us. Retrieved June 17, 2024, from https://www.unepfi.org/about/

UNEP-FI. (2023a). Nature-Positive Insurance: Evolving Thinking. Retrieved June 17, 2024, from https://www.unepfi.org/industries/insurance/nature-positive-insurance-evolving-thinking-and-practices/

UNEP-FI. (2023b). UN-convened Net-Zero Insurance Alliance Target-Setting Protocol. Retrieved June 17, 2024, from https://www.unepfi.org/wordpress/wp-content/uploads/2023/01/NZIA-Target-Setting-Protocol-Version-1.0.pdf

United Nations Department of Economic and Social Affairs. (2024). THE 17 GOALS. Retrieved May 10, 2024, from https://sdgs.un.org/goals

Verdantix. (2017). Sustainability Reporting Hits A Plateau As The Number Of Reports Using The GRI Guidelines Falls For The First Time. Retrieved June 6, 2024, from https://www.sustainability-reports.com/sustainability-reporting-hits-a-plateau-as-the-number-of-reports-using-the-gri-guidelines-falls-for-the-first-time/

Wagenhofer, A. (2024). Sustainability Reporting: A Financial Reporting Perspective. *Accounting in Europe*, (1), 1–13.

World Business Council for Sustainable Development and World Resources Institute. (2004). Retrieved May 24, 2024, from https://files.wri.org/d8/s3fs-public/pdf/ghg_protocol_2004.pdf

World Commission on Environment and Development. (1987). Our Common Future. Retrieved May 11, 2024, from https://www.are.admin.ch/are/en/home/media/publications/sustainable-development/brundtland-report.html

World Economic Forum. (2020). Measuring Stakeholder Capitalism: Towards Common Metrics and Consistent Reporting of Sustainable Value Creation. Retrieved June 25, 2024, from https://www3.weforum.org/docs/WEF_IBC_Measuring_Stakeholder_Capitalism_Report_2020.pdf