JUMS

Volume 6, Issue 1, March 2021

JUNIOR MANAGEMENT SCIENCE

Olivia Hohlwegler, The Value of CSR in Times of Increased Policy Uncertainty: Evidence from the Brexit Referendum 1

Maximilian Kühn, Are Firms Paying for the Minimum Wage? Evidence from Germany 25

Julia Vetter, The Glass Cliff – Women's Thrive to Save Poor Performance and how to Approach it in the Workplace 39

Kimberly Klebolte, Employment Protection Legislation, Youth Unemployment and the Role of the Educational System 60

Michael Amroudi, The Impact of Management, Family and Employee Ownership Concentration on Firm Performance 81

Daniel Dyck, Der Einfluss der Besteuerung auf Managementanreize und die Nutzung von Bonusbanken 100

Philipp Schiele, Modern Approaches to Dynamic Portfolio Optimization 149

Christina Winder, Analyse globaler Steuerstandards der OECD zu digitalen Geschäftsmodellen 190

Advisory Editorial Board:
DOMINIK VAN AAKEN
FREDERIK AHLEMANN
BASTIAN AMBERG
THOMAS BAHLINGER
CHRISTOPH BODE
ROLF BRÜHL
JOACHIM BÜSCHKEN
CATHERINE CLEOPHAS
RALF ELSAS
MATTHIAS FINK
DAVID FLORYSIAK
GUNTHER FRIEDL
FRANZ FUERST
WOLFGANG GÜTTEL
CHRISTIAN HOFMANN
KATJA HUTTER
LUTZ JOHANNING
STEPHAN KAISER
NADINE KAMMERLANDER
ALFRED KIESER
NATALIA KLIEWER
DODO ZU KNYPHAUSEN-AUFSEß
SABINE T. KÖSZEGI
ARJAN KOZICA
CHRISTIAN KOZIOL
TOBIAS KRETSCHMER
HANS-ULRICH KÜPPER
ANTON MEYER
MICHAEL MEYER
GORDON MÜLLER-SEITZ
J. PETER MURMANN
ANDREAS OSTERMAIER
BURKHARD PEDELL
MARCEL PROKOPCZUK
TANJA RABL
SASCHA RAITHEL
NICOLE RATZINGER-SAKEL
ASTRID REICHEL
KATJA ROST
THOMAS RUSSACK
MARKO SARSTEDT
ANDREAS G. SCHERER
STEFAN SCHMID
UTE SCHMIEL
CHRISTIAN SCHMITZ
PHILIPP SCHRECK
GEORG SCHREYÖGG
LARS SCHWEIZER
DAVID SEIDL
THORSTEN SELLHORN
ANDREAS SUCHANEK
ORESTIS TERZIDIS
ANJA TUSCHKE
STEPHAN WAGNER
BARBARA E. WEIßENBERGER
ISABELL M. WELPE
HANNES WINNER
THOMAS WRONA
THOMAS ZWICK

Published by Junior Management Science e.V.

Junior Management Science 6(1) (2021) 1-24

Junior Management Science

journal homepage: www.jums.academy

The Value of CSR in Times of Increased Policy Uncertainty: Evidence from the Brexit Referendum

Olivia Hohlwegler

University of St Andrews

Abstract

Scrutinising the 2016 Brexit Referendum, this paper examines the impact of economic policy uncertainty on the relationship between corporate financial performance and CSR using a sample of 320 non-financial firms listed at the London Stock Exchange. The sample covers the period of 2014 to 2018 with 2016 marking increased, and 2017 and 2018 representing years of moderated policy uncertainty. Using cross-sectional regression analysis of shock-period buy-and-hold returns, this paper finds (I) a statistically and economically significant inverse relationship between reservoirs of social capital previously accrued through CSR initiatives and returns. The effect is driven by the Governance component of CSR, whereas the impact of combined Environmental and Social pursues was not found to be meaningful. Using difference-in-difference methodology with continuous treatment, this investigation concludes (II) high-CSR and low-CSR firms do not differ in terms of sensitivity to adverse Brexit shock implications on operating performance, profitability, financial health, and firm value. Further, (III) effects of CSR on aforementioned aspects of real performance were not found to vary alongside levels of policy uncertainty.

Keywords: Brexit referendum; CSR; ESG investing; policy uncertainty.

1. Introduction

In the title of his article from 1970, Milton Friedman famously claimed that "The Social Responsibility of Business is to Increase Its Profits". Almost 40 years after one of the most influential neoclassical economists negated the fundamental question of whether firms should go beyond investing in activities with direct link to profit generation, corporations and investors draw a different picture: Over the last two decades, Corporate Social Responsibility (CSR) and environmental, social and governance (ESG) initiatives[1] have developed into popular buzzwords in business literature (Rüegg-Stürm, 2014). At the end of 2017, 12 trillion dollars were invested in socially responsible assets in the USA alone (Forum, US Social Investment, 2018). Likewise, 78% of the world's largest 4,900 companies disclose information on governance mechanisms and their social and environmental impact (KPMG, 2017). The justification behind pursuits of socially responsible activities are that they create a form of intangible, alternative capital: the so-called social capital comprising relational wealth among stakeholders. Important component of this capital is increasing trustworthiness of firms. Knack and Keefer (1997); Putnam (1993) argue this to reduce transaction costs as stakeholders attribute lower uncertainty about whether firms honour their contracts to firms they trust.

Conversely, the last decade of expanding trust-building initiatives to mitigate uncertainty among stakeholders was accompanied by numerous political events inducing policy uncertainty[2] across markets affecting stock markets and real performance (inter alia, Baker et al., 2016; Kelly, Pastor, &

[1] Both concepts of CSR and ESG (used interchangeably in this paper) refer to corporate activities beyond profit generation that include, amongst others, "civic engagement, shared beliefs, and disposition towards cooperation between the firm and its stakeholders" (Lins, Servaes, & Tamayo, 2017, p. 1790). Thereby, firms go above and beyond legal requirements, which might well be justified by strategic reasoning (Jo & Harjoto, 2012). The acronym CSR has its roots in managerial literature based on Ansoff (1965) and Freeman (1984) Stakeholder Approach to Strategic Management. In contrast, ESG is employed in the context of Socially Responsible Investing (SRI). Hence, the concept of ESG might be understood as a form of measurement of CSR (Bouslah, Kryzanowski, & M'Zali, 2013). Following authors such as Bouslah et al. (2013); Bouslah, Kryzanowski, and M'Zali (2018); Di Giuli and Kostovetsky (2014); Renneboog, Ter Horst, and Zhang (2008) this paper proxies CSR using all three ESG dimensions.

[2] Policy uncertainty is defined by Baker, Bloom, and Davis (2016) as uncertainty about economic policy decisions. They identify four elements of economic policy uncertainty: uncertainty regarding the (1) identity of deci-

DOI: https://doi.org/10.5282/jums/v6i1pp1-24

Veronesi, 2016; Pastor & Veronesi, 2012; Voth, 2002).

The coinciding trends raise the question as to whether firms can actively instrumentalise trust-building CSR initiatives to mitigate adverse effects of uncertainty on corporate financial performance. To the best knowledge of the author, no previous investigation was conducted to scrutinise the effects previously accumulated social capital has on corporate financial performance during times of increased policy uncertainty. Instead, literature has treated the relationship between financial performance and CSR, and financial performance and policy uncertainty as individual research streams.

As for CSR and firm performance, literature is abundant and has focused on establishing causality and directionality of effects of CSR on different aspects of corporate financial performance. Authors such as Brammer, Brooks, and Pavelin (2006); Jo and Na (2012); Kim, Li, and Li (2014) focus on stock returns and idiosyncratic market risk. Bird, D. Hall, Momentè, and Reggiani (2007); Guenster, Bauer, Derwall, and Koedijk (2011); Jiao (2010); Jo and Harjoto (2011) investigate operating performance and firm value. Also access to finance was addressed in literature (Cheng, Ioannou, & Serafeim, 2014; Cui, Jo, & Na, 2018; Ghoul, Guedhami, & Kim, 2017; Harjoto & Jo, 2015). Two opposing schools of thought have emerged explaining the relationship: the stakeholder view of the firm that promotes positive effects through reputation building (e.g. Jo & Harjoto, 2012; Makni, Francoeur, & Bellavance, 2009), and the shareholder view of the firm arguing in favour of Friedman based on overinvestment costs (e.g. Barnea & Rubin, 2010; Cespa & Cestone, 2007). The relative justification of one theory over the other has received much attention, but no consensus was reached so far.

As for policy uncertainty, stock market reactions to elections (Kelly et al., 2016; Li & Born, 2006; Pantzalis, Stangeland, & Turtle, 2000) and to Baker et al. (2016) economic policy uncertainty index were analysed. Further, studies by N. H. Nguyen and Phan (2017); Phan, Nguyen, Nguyen, and Hegde (2019); Xu, Chen, Xu, and Chan (2016) also investigate firm behaviour. But little research addresses differences in sensitivity to uncertainty in the cross-section of firms (Hill, Korczak, & Korczak, 2019).

More recent shock-based papers on CSR provide context for a link between policy uncertainty and CSR's effects on firms: Bouslah et al. (2018); Buchanan, Cao, and Chen (2018); Lins et al. (2017), focusing on the 2008 financial crisis, indicate the two opposing theories on CSR describe effects that exist relative to each other. Overall dominance of one effect over the other is influenced by macro-level conditions, with dominance being more pronounced under adverse conditions.

One of the most recent and prominent – since still persisting – source of uncertainty is the European Union Membership Referendum (Brexit Referendum) that took place on 23rd June 2016. The Referendum has proven that policy uncertainty is a likewise source for adverse macro conditions. The vote of UK citizens to leave the European Union (EU) has led policyrelated uncertainty in the UK soar to unprecedented heights. At the same time, the UK's economy is confronted with unforeseeable consequences. The exit from the EU planned for October 2019 will overhaul the overall legal framework within which companies interact in the UK and across UK borders (Edgington, 2019; Hill et al., 2019).

This paper advances both research streams on CSR and policy uncertainty by unifying them to analyse the impact of policy uncertainty on the relationship between CSR and corporate financial performance. As both research streams are prone to the endogeneity problem (Harjoto & Jo, 2015; Julio & Yook, 2012; Ullman, 1985), we apply a shock-based causal inference methodology. Thereby, the Brexit Referendum provides the natural experiment to close the identified research gap, using a sample of 320 non-financial firms listed at the London Stock Exchange over the period of 2014 to 2018. We further progress literature by disentangeling contributions of overall CSR by analysing effects of the combined Environmental and Social, as well as the Governance pillar of CSR. A third contribution is to compare evaluations of CSR by investing stakeholders (by analysing buy-and-hold returns in an event study) with evaluations by non-investing stakeholders (by analysing operation performance, financial health, and firm value using difference-in-difference analysis).

The following conclusions have been drawn. CSR investments, especially Governance related pursues, can destroy shareholder wealth during policy uncertainty. However, from the view of non-investing stakeholders, no effects on real performance are evident – for none of the analysed aspect of CSR. Further, the latter relationship does not vary across different levels of policy uncertainty. The findings are robust across a series of robustness tests.

These findings have various implications for practitioners. For investors, they suggest avoiding high-CSR firms during increased policy uncertainty, especially if they have established Good Corporate Governance practices. From the perspective of financial managers, we infer CSR investments are not an effective means to limit exposure to policy uncertainty shocks.

Our findings are also relevant for politics given that the US Congress is debating on mandatory ESG disclosure regulations (Harty, 2019), and given that von der Leyen, the newlyelected president of the EU Commission, addressed plans of environmental taxes in her campaign speech (Stearns, 2019). Negative effects on stock returns and neutral effects on real performance suggest ESG dimensions are still areas in which Government has to intervene for companies to meet legitimate concerns of certain stakeholders.

The remainder of this study is structured as follows. Section 2 discusses the theoretical motivation behind the effect of social capital on corporate firm performance and the link to

sion makers, the (2) nature and (3) timing of policy actions, and economic ramifications of "noneconomic" policy matters such as military actions. Baker et al. (2016) policy uncertainty index has established as a generally accepted measure in financial research (see, inter alia, Fernández-Villaverde, Guerrón-Quintana, Kuester, and Rubio-Ramírez (2015); N. H. Nguyen and Phan (2017); Pastor and Veronesi (2012, 2013)). Thus, Baker et al. (2016) definition is followed in this paper.

policy uncertainty. Section 3 discusses the data analysed and motivates the natural experiment employed for regression analysis. Section 4 and 5 present and discuss results of baseline and robustness tests. Section 6 discusses the paper's findings in context with previous shock-based inferences. Section 7 concludes.

2. Theoretical Framework

2.1. Social Capital and Corporate Financial Performance

The question of main interest in financial research on CSR links back to Friedman's article from 1970: Can companies' socially responsible behaviour be brought in line with investors' interest of value maximisation? Two schools of thought have developed in literature offering opposing answers to this question: (I) the stakeholder, and (II) the shareholder view of the firm.

The stakeholder view has its roots in Ansoff (1965) and Freeman (1984) seminal works and suggests a positive relationship between CSR activities and corporate financial performance. The argument supporting the positive effect is the "reputation-building hypothesis" (Jo & Harjoto, 2012; Makni et al., 2009). Its argument is that social, ethical, and environmental as well as good corporate governance initiatives build up a reputation among external parties in the form of goodwill and trust in the company, i.e., create so-called social capital[3] (Guiso et al., 2004, 2008; Sacconi & Degli Antoni, 2011). More specifically, Luo and Bhattacharya (2009) highlight the value of CSR to firms stems from the beliefs stakeholders hold about the firm. If CSR activities and conveyed implications about the firm receive positive evaluation from stakeholders, social capital is generated. Social capital, in turn, is hypothesised to increase financial performance by generating different forms of relational wealth among different stakeholder groups (Godfrey, 2005).

For non-investing stakeholders, Lins et al. (2017) mainly trace positive effects back to creating moral incentives for stakeholders to honour their contract. Given the information asymmetries between the parties and the fact that contracts are on future transactions, stakeholders and companies always face the risk of the other party not honouring them. From the stakeholders' perspective, CSR activities are theorised by Putnam (1993) to foster trust and cooperation and to mitigate conflicts (Freeman, 1984). This argumentation is supported by empirical evidence of a positive relationship between perceived trustworthiness of firms and stakeholder cooperation (see, inter alia, Guiso, Sapienza, & Zingales, 2015; Servaes & Tamayo, 2013). Knack and Keefer (1997) infer stakeholders may perceive the breaching of contracts by the firm as less likely, which reduces the need for formal contracts and thus transaction costs. In this context, Sharfman and Fernando (2008) and El Ghoul, Guedhami, Kwok, and Mishra (2011) find evidence for high CSR reducing the cost of equity capital and Goss and Roberts (2011) for reduced default risk.

As for shareholders, Guiso et al. (2008) explain that due to agency problems, the investment decision is not only driven by accounting data, but also by an act of faith in the reliability of disclosed information. Building on this, Lins et al. (2017) also argue trust-based measures need to be considered. Therefore, social capital measures are requested to provide additional information on the company's integrity and credibility, which leads to higher relative valuation in cases of positive perception (Bouslah et al., 2013; Lee & Faff, 2009). In this sense, social capital can be understood as a means of reducing information asymmetries, which is empirically supported by Cui et al. (2018). Ghoul et al. (2017) add the effect is especially pronounced in countries that lack market-supporting institutions.

An opposing negative effect of CSR on corporate financial performance is put forth by the second school of thought, the shareholder view of the firm. This school does not dispute positive effects of social capital's relational wealth. Rather, a managerial opportunism argument is derived from the agency theory, the so-called "overinvestment hypothesis". This argument follows Friedman (1970) notion that costs of CSR might outweigh the benefits. According to the agency theory, shareholders have fewer insights on business operations than managers due to the separation of ownership and control (Shleifer & Vishny, 1997). In case of diverging interests, Buchanan et al. (2018) theorise managers can exploit the incomplete contracts to pursue their interests and potentially harm investors in their wealth maximisation. Scholars supporting the shareholder view posit managers do have an intrinsic interest in investing in CSR activities beyond an optimal, shareholder-benefitting level; either to increase their reputation as socially responsible managers (Barnea & Rubin, 2010) or to reduce the probability of replacement (Cespa & Cestone, 2007). Even though Eccles, Ioannou, and Serafeim (2014) argue CSR can reduce agency problems since managers are forced to apply a long-term focus and are thus associated with more responsible decision-making, Surroca and Tribó (2008) find empirical evidence for managers exploiting CSR for opportunistic pursuits. Their findings suggest entrenched managers employ CSR investments as a defence strategy. In this sense, CSR investments might diverge company resources and thus harm real performance. Orlitzky (2013) posits overinvestments also have a negative effect on stock returns. He suggests the created social capital is not systematically correlated with financial fundamentals of the firm. The results are unjustified excess market valuations leading to excess market volatility.

An abundant stream of literature is devoted to proving which theory holds in reality. Table 7 in the appendix pro-

[3] The concept of social capital itself stems from the field of sociology. In this context, La Porta, Lopez-de Silanes, Shleifer, and Vishny (1997, p. 333) define it as "a propensity of people in a society to cooperate to produce socially efficient outcomes". The concept was operationalised by economists such as Coleman (1988) or Guiso, Sapienza, and Zingales (2004, 2008) to be understood as a resource for profit-generating activities. In an OECD paper, Scrivens and Smith (2013) decompose this resource into four pillars: relationships, social network support, civic engagement, and trust. This definition overlaps with the concept of CSR and ESG outlined in footnote 1. For a detailed theoretic motivation of the concepts' link, see Sacconi and Degli Antoni (2011).

vides an examplary overview and hints at the conclusion that literature is inconclusive. Waddock and Graves (1997) attribute this to inconsistencies in conceptualising CSR as different studies focus on different components of CSR (see Table 7). More severely, Ullman (1985) criticises that problems of endogeneity and the omitted variable bias are another explanation for inconsistencies, discrediting the validity of past research.

The methodological issue raised by Ullman (1985) is attempted to be overcome by more recent, shock-based studies which offer a different conclusion: In previous research designs, the relationship is mostly understood to be time-invariant. Buchanan et al. (2018, p. 76), however, propose the two school of thoughts are less of two mutually exclusive explanations of which one is theorised to be correct and the other to be wrong. Rather, they have to be understood as antagonising channels through which CSR takes effect and which exist alongside each other. They suggest the relationship is time-varying and the overall effect depends on which mechanism is dominant during the time period in question.

Based on the stakeholder theory, Lins et al. (2017) derive from the notion of reciprocity of social capital that in difficult times, stakeholders are more likely to "do whatever it takes" to support firms with high social capital. As examples, they name employees working harder, and outside agencies being more sympathetic and supportive. In contrast, Buchanan et al. (2018) argue that under the shareholder theory, shocks are expected to increase overinvestment effects.

As for idiosyncratic, firm-specific adverse legal events, Godfrey, Merrill, and Hansen (2009) find evidence for social capital acting as an insurance policy. They find that after the event, high CSR scores have a positive effect on two-day cumulative abnormal returns. Similarly, Minor (2015) suggests firms with high CSR investments can, on average, save $1 billion in firm value upon negative legal events. More recent and more closely related papers focus on aggregated market shocks. Lins et al. (2017) investigate three adverse events: the 2008 financial crisis, the Enron, and the Worldcom scandal. All three events led to a decrease in the aggregated level of trust in companies in the US. For the first two events, they find a positive relationship between CSR and stock returns; for the first event they also find a positive relationship with operational performance. They conclude that social capital especially pays off in crises of trust. Two more studies analyse the risk-mitigation effects over varying economic conditions, both two focusing on the 2008 financial crisis. Bouslah et al. (2018) investigate the effect of social performance on idiosyncratic risk. Their results suggest that the relationship significantly depends on the macroeconomic condition, with an effect of reducing volatility by 1.18% to 1.84% during the crisis period. Buchanan et al. (2018) come to an opposing conclusion. Even though they observe a higher firm value as measured by Tobin's Q for CSR firms before the crisis, the opposite holds for the crisis period. This might seem contradictory given Lins et al. (2017) finding of a positive impact on short-term operating profitability. However, Tobin's Q is a more forward-looking market valuation than accounting-based returns and profitability measures (Guenster et al., 2011). Hence, the two findings might indicate that even though firms with a high reservoir of social capital are assigned to a valuation premium and are less affected in short-term profitability, the market does not perceive high CSR firms to outperform in terms of future cash flows when faced with a market downturn.

2.2. Corporate Financial Performance in Times of Policy Uncertainty

Findings of Lins et al. (2017) and Buchanan et al. (2018) indicate that the effect of social capital on corporate firm performance is time-varying, depending on the market condition; in their case predominantly analysed around the 2008 financial crisis. Building on Guiso et al. (2008, p. 2557) argument that making business with firms is an act of faith in the reliability of the firm's information disclosure, Lins et al. (2017, p. 1791) argue that the trust element of social capital should pay off more during economic turmoil. This paper posits an increase in policy uncertainty is an event negatively affecting firms that allows investigating whether previously accrued social capital helps firms to weather difficult times. In fact, the effects of policy uncertainty on firm performance have been investigated extensively. And in fact, an abundant research stream has emerged emphasising adverse effects on financial markets and the real economy.

One well-established research stream is devoted to analysing the impact of uncertaintybearing recurring elections, termed electoral uncertainty by Kelly et al. (2016). Several studies focus on its impact on the aggregated stock market to analyse the existence of risk premia for electoral uncertainty. Abnormally high stock returns shortly before elections were observed by Pantzalis et al. (2000) for national elections in 33 countries and by Li and Born (2006) for US presidential elections. Together with Gao and Qi (2013) finding of rising municipal bond yields accompanying US gubernatorial elections, these findings are indicative of an existing risk premium, even though Białkowski, Gottschalk, and Wisniewski (2008) could not support this across 27 markets. In a similar vein, stock market volatility around elections was analysed. Findings by Kelly et al. (2016) across 271 international elections and G20 summits, by Gemmill (1992) for the 1987 British parliamentary election, and by Goodell and Vähämaa (2013) for US presidential elections suggest an increase in implied volatility. Boutchkova, Doshi, Durnev, and Molchanov (2012) offer more granularity: industries which are more sensitive to politics experience higher return volatility. In contrast to these aggregated effects, Julio and Yook (2012) and Jens (2017) study real effects of electoral uncertainty. Both studies suggest that firms reduce investments before US national and gubernatorial elections.

N. H. Nguyen and Phan (2017) and Pastor and Veronesi (2012) indicate these studies only capture part of policy uncertainty – a broader concept than electoral uncertainty. Following the definition of Baker et al. (2016), the first concept also refers to policy changes affecting the macroeconomic setting such as trade agreements and tax legislation, as

well as economic impacts of non-economic political events. This broader concept has been analysed regarding macroeconomic effects by Baker et al. (2016) and Fernández-Villaverde et al. (2015). Both reason negative impacts on overall economic conditions. This was for instance, illustrated by Standard & Poor's first-ever downgrading of US Treasury bonds due to policy uncertainty (Kelly et al., 2016). Similarly, Stock and Watson (2012) indicate policy uncertainty negatively affected recovery after the financial crisis. This notion is further supported by Giavazzi and McMahon (2012) linking policy uncertainty to reductions in household consumption and savings increases. As for influences on aggregated stock markets, academic literature follows studies on electoral uncertainty investigating a potential risk premium for political uncertainty measured by Baker et al. (2016) index or the International Country Risk Guide (see, inter alia, Erb, Harvey, & Viskanta, 1996). This notion is supported by Voth (2002) finding of increased stock volatility. Another researched aspect is company behaviour. According to Gulen and Ion (2016), companies become more selective; especially firms with higher degrees of investment irreversibility and higher financial constraints. This is further supported by Bloom, Bond, and van Reenen (2007) and Jens (2017). Another set of studies is dedicated to real effects on corporate financial performance. In terms of profitability, Belo, Gala, and Li (2013) establish that cash flows vary over political cycles in the US. This is in line with Cremers and Yan (2016) suggestion of uncertainty increasing cash flow volatility and ultimately default risk. Similarly, An, Chen, Luo, and Zhang (2016) observe that cash flow volatility caused by changes in investment leads to reduced profitability as measured by return on assets in China. Together with Phan et al. (2019) and Xu et al. (2016) linking uncertainty to reduced cash holdings, one might infer increased risk of financial distress. Investigations of the cost of capital provide further support. Waisman, Ye, and Zhu (2015) found bond spreads to increase through uncertainty. Francis, Hasan, and Zhu (2014) support this increase in debt financing costs by detecting an additional risk premium for firms' exposure to policy uncertainty priced in bank loan spreads.

Despite the event's topicality, some studies analyse the Brexit Referendum's impact on economy. Hill et al. (2019) subsume the Brexit Referendum to be a shock to policy uncertainty as it has consequences which are more far-reaching than most parliamentary or presidential elections. As for Brexit, the debate does not focus on selected policies but rather affects the total of government policies that shape the business environment. The vote thus leaves companies facing uncertainty about future operations in the UK. For instance, international trading agreements, immigration policies and the overall legal framework primarly driven by EU law are to be overhauled. Despite anticipated impact on corporate decision making, Davies and Studnicka (2018) comment most research addresses macroeconomic consequences. Even though Minford, Gupta, Le, Mahambare, and Xu (2016) expect positive impacts on the UK[4], most studies suggest severe negative effects such as a decline in foreign direct investment (see, inter alia, Dhingra, Ottaviaono, Sampson, & van Reenen, 2016; Fraser of Allander Institute, 2016). Focusing on the stock market, Davies and Studnicka (2018) detect significant negative abnormal returns for FTSE350 companies, especially for firms with global value chains. This is in line with Boutchkova et al. (2012) but contradicts Beaulieu, Cosset, and Essaddam (2005). In contrast to the various studies on uncertainty on an aggregated level, several studies on Brexit analyse how different firms and industries are affected. Tielmann and Schiereck (2017), for instance, report a negative valuation effect for the logistic sector, while Ramiah, Pham, and Moosa (2017) suggest the banking industry is especially strongly affected. The latter finding is supported by Hill et al. (2019). They further find that internationalisation moderates and greater firm size magnifies effects of Brexit uncertainty. Further, growth firms, due to their dependence on further investments, are especially hard-hit. However, to the best knowledge of the author, no paper has investigated the influence of CSR on the exposure to policy uncertainty.

2.3. CSR, Firm Performance, and Policy Uncertainty

The literature review on the relationship between CSR and policy uncertainty respectively on corporate financial performance reveals an overlapping research gap in both streams of literature. As for policy uncertainty, aggregated adverse effects on the stock markets, firm behaviour, and operating performance are well documented. But even though policy uncertainty has reached unprecedented hights (see Figure 1), "we know surprisingly little about how this uncertainty affects different firms and industries" (Hill et al., 2019, p. 58). While the few recent studies on Brexit address this gap by analysing individual sectors and different firm characteristics such as internationalisation, to the best knowledge of the author, no investigation focusing on whether accrued social capital mitigates effects of uncertainty has been conducted so far. This leads us to the corresponding gap in CSR literature. Three papers have indicated that the overall effect of CSR depends on the relative, time-varying dominance of the reputation-building benefits over the overinvestment costs. The investigations establish the overall dominance is influenced by aggregated macro-conditions. However, the studies mainly focus on the 2008 financial crisis, with the USA being the only market analysed. Hence, it still has to be established whether findings also hold during other adverse macro-events and in different markets.

In this sense, this paper scrutinises the 2016 Brexit Referendum to contribute to closing both research gaps. By analysing how previously accrued social capital influences financial performance of firms listed at the London Stock Exchange in the year of and the two years after the Referendum (moderated policy uncertainty), this paper advances the

[4] Sampson, Dhingra, Ottaviaono, and van Reenen (2016) criticise underlying assumptions to be too optimistic.

literature on CSR as well as on policy uncertainty. We further combine several research streams by scrutinising both stock market and real performance of firms. Real performance is captured on the level of operating performance, financial health, and firm value. This not only captures a broad picture of financial performance but also differences in how investing and non-investing stakeholders value CSR. Following Buchanan et al. (2018); Lins, Volpin, and Wagner (2013), we investigate CSR effects on corporate financial performance during the shock and the recovery period to scrutinise whether effects are unique to times of high uncertainty.

In terms of stock market performance, literature on policy uncertainty indicates the existence of risk premia required by investors (inter alia, Erb et al., 1996; Voth, 2002). Underlying causes might be increased cash flow volatility and default risk (An et al., 2016; Cremers & Yan, 2016). If the stakeholder view holds, i.e. if CSR has positive effects on returns, then shareholders' positive valuation of integrity and credibility should cause investors to be less pessimistic about future firm performance. This, in turn, should lead to valuation premia for high CSR firms (Godfrey, 2005; Guiso et al., 2008; Lee & Faff, 2009). This leaves us with the following first hypothesis:

> H_0^1: Social capital is positively related to shock-period stock returns.

In terms of real performance, the stakeholder view posits that the relational wealth created among different stakeholders – e.g. employee commitment, customer loyalty, credibility among investors and lenders – acts as a reservoir of goodwill. This is related to the notion of reciprocity suggesting that during difficult times stakeholders which have been accommodated in the past are willing to "do whatever it takes" to support firms in need. This can relate back to different levels of firm performance during policy uncertainty. In terms of real performance, the stakeholder view posits that the relational wealth created among different stakeholders – e.g. employee commitment, customer loyality, credibility among investors and lenders – acts as a reservoir of goodwill. In times of weakened economic conditions, customers might, for instance, be more willing to "stick" with companies of which they hold higher beliefs and lenders could be more complaisant (Lins et al., 2017). Further, previouslybuilt trust could dampen increases of information asymmetries in times of uncertainty (Cremers & Yan, 2016). Following the stakeholder approach, these considerations might shield firms from reduced profitability due to policy uncertainty (An et al., 2016). These considerations leave us with the following two hypotheses for operational performance (measured by gross marging, sales growth, return on assets, and return on equity), and financial health (measured as financial slack and Altman's Z):

> H_0^2: Social capital is more positively related to operating performance during policy uncertainty.

> H_0^3: Social capital is more positively related to financial health during policy uncertainty.

We include financial slack due to Xu et al. (2016) and Phan et al. (2019) findings of reduced cash holding and higher costs of lending during policy uncertainty (Francis et al., 2014). It can be interpreted as a cushion of cash firms can use to weather economic downturns. This might be attributable to Eccles et al. (2014) argument of managers of high-ESG firms being more responsible.

Given the theorised positive effects that trust created through CSR has on operating performance and financial health, we anticipate the market to have more optimistic views on cash flow volatility caused by policy uncertainty. Eventually, this should positively impact firm value measured as Tobin's Q:

> H_0^4: Social capital is more positively related to firm value during policy uncertainty.

A fith and last hypothesis to be tested addresses another limitation of the shock-based CSR research designs. Table 7 indicates that different elements of CSR might be perceived differently by various stakeholder groups. For instance, Harjoto and Jo (2015) found that high overall CSR scores capturing legal and normative aspects increase firm value, whereas disaggregated high normative CSR scores are significantly linked to reduced firm value. In contrast, Bird et al. (2007) found a positive relationship on stock returns for employee-related activities, but an inverse relationship for environmental activities. Hence, it might well be anticipated that the different pillars of CSR affect corporate financial performance differently. CSR is disaggregated into two components: a combined Environmental and Social (ES),[5] and an individual Governance score.

> H_0^5: Different components of CSR have different effects on corporate financial performance.

In summary, this paper posits that in high-uncertainty periods, trust- and cooperative-based relationships between firms and stakeholders should pay off more than in more certain times. In this sense, we anticipate to find evidence in favour of the stakeholder view of the firm.

3. Research Design

3.1. Empirical context

Buchanan et al. (2018) point out research on the relationship between corporate financial performance and CSR faces to major challenges. The first is that the shareholder and stakeholder view constitute two antagonising channels for CSR effects with time-varying relative dominance. The second challenge is the directionality of the relationship, i.e. that CSR might not be endogenous to firm performance. For

[5] The rationale behind this is mainly based on the high correlation between the E and S scores. See section 4.1. for further explanation.

instance, Waddock and Graves (1997) find effects in both directions of influence and Hong, Kubik, and Scheinkman (2012) show increases in firms' financial constraints, on average, have a negative influence on CSR involvement. Motivated by scholars such as Buchanan et al. (2018); Derrien and Kecskés (2013); Duchin, Ozbas, and Sensoy (2010); Lins et al. (2017, 2013)[6], these challenges are addressed by applying a shock-based study design. Following Lins et al. (2017), this paper applies two shock-based investigations: For one, uncertainty-period buy-and-hold shock-period returns are analysed while difference-in-difference (DiD) methodology is applied to investigate financial real performance. Contrary to previous studies, a political shock, the Brexit Referendum on 23rd July 2016, is analysed.

3.2. Brexit Referendum as a Policy Uncertainty Shock

The conclusion of the Referendum being an exogenous shock was laid out by previous studies highlighting the multifaceted legal, macro-, and microeconomic effects of Brexit, which are still unforseeable in the year of the UK's permanent exit from the EU. This notion is supported by Figure 1 plotting Baker, Bloom, and Davis (2019) Policy Uncertainty Index. June 2016 marks the highest uncertainty level ever recorded. Additionally, Brexit could not have been anticipated in advance as Brexit polls in the one-and-a-half years leading to the vote (FIGURE 2) indicate neither the Leave nor the Remain side had gained a persistent lead. Only in the few months prior to 23rd of June, the Remain party had gained a slight majority and even despite this, the opposite outcome was achieved. Both notions hint at the event meeting Atanasov and Black (2016, p. 216) requirement of an exogenous shock with permanent effect for shock-based causal inference studies such as event studies and DiD.

The second peak of policy uncertainty in 2016 might raise the concern that the Referendum might not have marked an isolated shock, violating a second DiD credibility criterion by Atanasov and Black (2016, p. 216). However, a look at the Brexit timeline reveals the spike coincides with political reactions to the Referendum. For instance, the First Minister of Scotland announced the intention to pass a Scottish independence referendum bill on 13 October 2016 given that the majority of Scots voted Remain (Walker, 2019).

A glance at current strategic decisions of corporations reveals a picture beyond estimated effects on macroeconomic conditions and consequences of Brexit procedures on stock markets as analysed in literature so far. Rather, it becomes apparent that corporations increasingly adapt their strategies to be prepared to deal with possible Brexit scenarios. British Steel has announced 400 job cuts in September 2018 because of the weakened pound (Sommerlad & Chapman, 2019). Similarly, Barclays has moved £190bn of assets under management to Ireland to avoid losing passporting rights allowing to serve customers EU-wide, with HSBC, RBS, and Lloyds following their lead (Crow, 2019). Turing to non-financial industry, a 2019 survey by the Institute of Directors surveying 1,200 business leaders suggests that ca. 30% of the surveyed either already have moved or are actively considering moving operations outside the UK (BBC, 2019). Given the various actions by firms to prepare for the worst-case scenario of a no-deal Brexit, it can be assumed that the shock is also sufficiently strong for DiD analysis (see Atanasov & Black, 2016, p. 216).

In summary, overall evidence suggests the Brexit Referendum to be a suitable event for shock-based causal inference study.

3.3. Sample Construction

The sample construction is based on financial and CSR data available on Thomson Reuter's databases Datastream and ASSET4, respectively. ASSET4 was introduced in 2002 and has been widely used in literature to investigate the effects of CSR on financial performance (see, inter alia, Cheng et al., 2014; Ghoul et al., 2017). The database rates companies across environmental, social, and governance aspects[7] on an annual basis. For the analysis of the individual pillars of CSR, the aggregated Environmental and Social, and the individual Governance scores provided by ASSET4 on Datastream were used. For the aggregated CSR analysis, a composite CSR score was calculated as the equally-weighted average of the E, S, and G scores (see, inter alia, Cheng et al., 2014).

ASSET4 comprises a worldwide universe of 7,975 historical constituents. Among those, 566 are public companies listed in the UK. To investigate whether accumulating social capital is beneficial for financial performance in uncertain times, the effects of pre-Brexit ESG scores on financial performance after the referendum are analysed. Even though the scores as of 2015 would have been the most recent scores available prior to the referendum, this paper relies on the scores as of 2014. This is done to control for two antagonising effects documented in literature: Buchanan et al. (2018); Lins et al. (2017) indicate firms might increase CSR investments as a risk management tool if adverse events are anticipated. Given that the people's vote was one of the conservative party's campaign pledges in the 2015 general election (Wheeler, 2017), some firms might have anticipated an increase in policy uncertainty. On the contrary, Julio and Yook (2012) provide evidence for the tendency that firms reduce investments in election years. The necessity to do so is illustrated by the high variance between ESG scores of 2014 and 2016. Based on these considerations, the ASSET4 universe as of 2014 comprising 405 companies constitutes the initial

[6] All these studies investigate whether firms with different characteristics (e.g. corporate governance or CSR practices) are affected differently by the 2008 financial crisis.

[7] Thomson One makes use of more than 400 different data points to construct 178 comparable measures, from which scores for ten categories are calculated. These categories are aggregated to form the three dimensions of Environmental (categories: resource use, emissions, innovation), Social (categories: workforce, human rights, community, product responsibility), and Governance (categories: management, shareholders, CSR strategy) scores (Reuters, 2019).

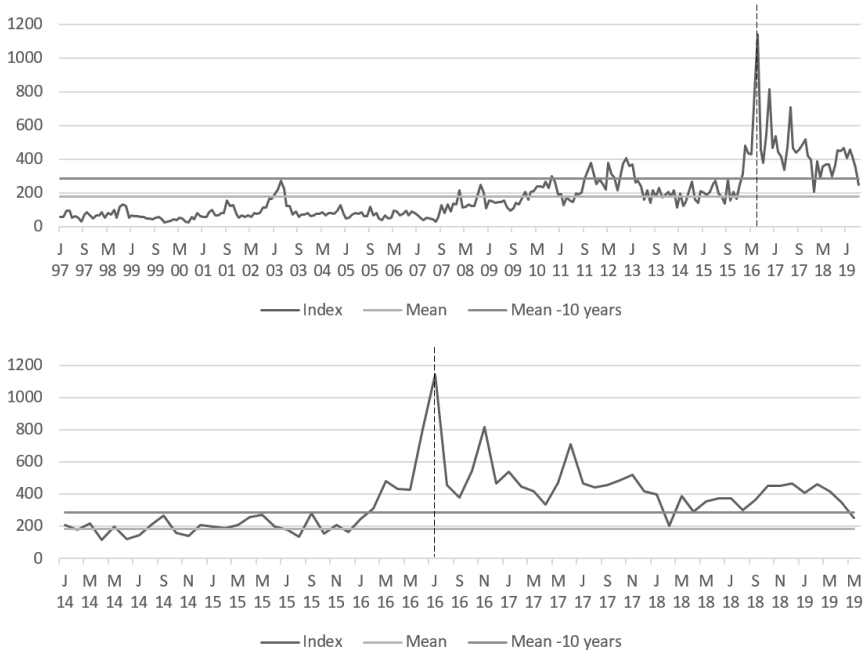


Figure 1: UK News-Based Policy Uncertainty Index (data by Baker et al. (2019))

The first graph shows the development of the UK Policy Uncertainty Index since the beginning of records; the second graph shows the development for the sample period of the data analysis. The line labelled mean shows the mean Policy Uncertainty Index since the beginning of records; the line labelled mean -10 years shows the mean for the last ten years. The dotted line indicates the date of the Brexit Referendum.

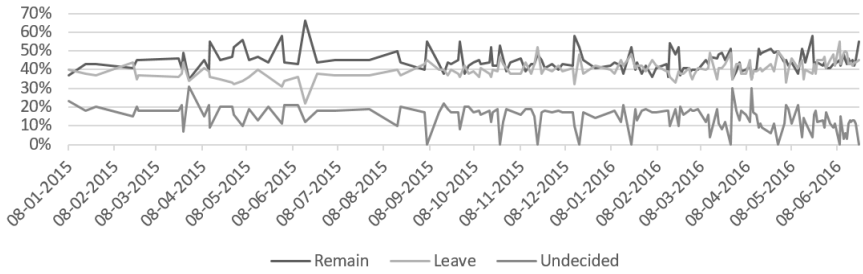


Figure 2: Outcome of Brexit Polls Conducted in 2015 and 2016 (based on Financial Times, 2016)

sample[8].

Following general convention in literature, financial firms[9] were removed. For one, due to differences in balance sheet structure. Secondly, because Ramiah et al. (2017) and Hill et al. (2019) found UK financial firms to be more severely affected by Brexit, possibly due to the threat of losing their EU bank license passport. This leaves us with an unbalanced panel of yearly financial data for 320 firms, starting in 2014 and ending in 2018 (1,600 firm-year observations). The financial and CSR variables used are defined below.

[8]The information on the universe of ASSET4's historical constituents was obtained from Datastream.

[9]Financial firms were identified using the two-digit ICB industry codes provided by FTSE International and obtained from Datastream.

3.4. Empirical Methodology

3.4.1. Shock-Period Returns

First question of interest is the effect of CSR on stock returns during the shock period. Following Lins et al. (2017), shock-period stock returns are measured as the firm's raw buy-and-hold returns (BHAR). BHARs were chosen to allow for compounding effects and thus to better capture investors' tendency to buy and hold equities for multiple months under the existence of transaction costs (Brooks, 2014; Dutta & Jog, 2009). The holding period commences three months prior to and liquidation of the position is three months after the Referendum, resulting in a holding period from March to September 2016. As illustrated in eqn. (1), BHARs are calculated as the difference between the compounded shock-period returns of the firm and of a benchmark. Thereby, the UK market portfolio provided by the University of Exeter's Xfi

Table 1: Definition of Variables

Variable	Measure
Panel A: Cross-Sectional Regression	
ltd_atw	Long-term debt over total assets as of 31/12/2015.
dlc_atw	Short-term debt over total assets as of 31/12/2015.
cash_taw	Cash and short-term investments over total assets as of 31/12/2015.
sizew	Natural logarithm of total assets as of 31/12/2015 to account for the highly skewed nature of this variable.
beta_rmrfw	Exposure to market risk estimated over the period between 01/2014 to 12/2015.
beta_umdw	Exposure to momentum factor estimated over the period between 01/2014 to 12/2015.
beta_hmlw	Exposure to value factor estimated over the period between 01/2014 to 12/2015.
beta_smbw	Exposure to size factor estimated over the period between 01/2014 to 12/2015.
Panel B: DiD Model	
Dependent Variables	
tqw	We follow Buchanan et al. (2018) and measure firm value using Tobin's Q approximated as follows: $$tq_t = \frac{at_t + (chso * prcc_f_t) - ceq_t}{at_t}$$ at is total assets, chso is common shares outstanding, prcc_f is year-end stock price, and ceq is common equity. A value below 1 indicates a firm is undervalued; a value above 1 indicates it is overvalued.
sales_gw	Sales growth defined as current annual sales divided by sales of the previous year.
gmw	We follow Lins et al. (2017) and calculate gross margin as: $$gm_t = \frac{sales_t - cogs_t}{sales_t}$$ cogs is cost of goods sold.
zscorew	Altman (1993) Z-score. We follow Bouslah et al. (2018) and calculate as follows: $$zscore_t = 1.2 * \left(\frac{wcap_t}{at_t}\right) + 1.4 * \left(\frac{re_t}{at_t}\right) + 3.3 * \left(\frac{ebit_t}{at_t}\right) + \left(\frac{sale_t}{at_t}\right) + 0.6 * \left(\frac{mkval_t}{lt_t}\right)$$ wcap is working capital, re is retained earnings, ebit is earnings before interest and taxes, mkval is market value of equity, and lt is book value of total liabilities. The lower the zscore, the higher the likelihood of default.
s_cashw	Financial slack which measures the cash available to finance new projects. It is calculated as: $$s_cash_t = \frac{oancf_t - dpc_t + xrd_t}{at_t}$$ oancf is the net cash flow from operating activities, dpc is depreciation, and xrd is R&D expenses.
roew	Return on equity measured as net income in year t over total common equity in year t.
sales_pew	Sales in year t over employees in year t.
roaw	Return on assets measured as net income in t over total assets in t.

(Continued)

Table 1—continued

Independent Variables	
esg2014	Composite measure of CSR (ESG score) as of 2014.
envs2014	Environmental score as of 2014.
socs2014	Social score as of 2014.
es2014	Composite environmental (E) and social (S) score as of 2014.
gov2014	Corporate Governance score as of 2014.
sizew	Firm size measured as natural logarithm of total assets in t.
capexpw	Current capital expenditures standardised by current total assets.
fabaw	Tangibility proxied by current property, plant and equipment over current total assets.
cash_taw	Cash holdings measured by current cash and short-term investments over current total assets.
bleveragew	Book leverage as the ratio of current long- and short-term debt to current total assets.
turnoverw	We follow Ding, Ferreira, and Wongchoti (2016) and include the natural logarithm of average monthly volume over shares outstanding at the end of each year t in models on Tobin's Q and ROE.
rdw	R&D intensity measured as current research and development expenses over current sales.
advertw	Servaes and Tamayo (2013) show the effect of CSR on firm value depends on customer awareness measured by advertising expenditures over sales. Following Arora and Dharwadkar (2011), current selling, general and administrative expenses over current sales is used instead due to data availability.
Interaction Terms	
ibresg	Interaction term between the Brexit Referendum dummy (1 in 2016; zero otherwise) and esg2014.
ipostesg	Interaction term between the post Brexit Referendum dummy (1 in 2017 and 2018; zero otherwise) and esg2014.
ibres	Interaction term between the Brexit Referendum dummy (1 in 2016; zero otherwise) and es2014.
ipostes	Interaction term between the Brexit Referendum dummy (1 in 2017 and 2018; zero otherwise) and es2014.
ibrcg	Interaction term between the Brexit Referendum dummy (1 in 2016; zero otherwise) and gov2014.
ipostcg	Interaction term between the post Brexit Referendum dummy (1 in 2017 and 2018; zero otherwise) and gov2014.

Note: the suffix w denotes winsorization at the 1% and 99% level. To ensure a robust sample size, we follow Buchanan et al. (2018) and set R&D equal to zero when R&D expenses are missing.

Centre for Finance & Investment was used as a benchmark[10].

$$\mathrm{BHAR}_{i,t} = \prod_{t=1}^{t}\left[1+R_{i,t}\right] - \prod_{t=1}^{t}\left[1+\mathrm{E}\left(R_{M,t}\right)\right] \tag{1}$$

$R_{i,t} = \dfrac{P_{i,t}-P_{i,t-1}}{P_{i,t-1}}$

$\mathrm{E}\left(R_{M,t}\right)$ expected return of the market portfolio at time t

The BHARs are then analysed in a cross-sectional regression to determine whether and how they are influenced by ESG, ES, and Corporate Governance. The baseline model includes linear ESG, ES and Governance scores as continuous variables. To control for omitted variables that happen to be correlated with CSR measures and influencing inferences drawn, we follow Lins et al. (2017) and include three sets of controls. The first set is firm's financial health. Earlier studies suggest that policy uncertainty is inversely related to cash holdings (inter alia, Phan et al., 2019; Xu et al., 2016) and frim profitability (An et al., 2016) as well as positively related to costs of debt financing (Francis et al., 2014). Given that these circumstances increase the risk of financial distress (Cremers & Yan, 2016), one might assume that profitable and cash-rich firms performed better during the shock period. This effect is captured by the financial health variables. The second effect controlled for are firm-specific characteristics, which were found to forecast stock returns successfully (inter alia, Kong, Rapach, Strauss, & Zhou, 2011; Rapach & Zhou, 2013). Finally, we include the Fama-French-Carhartt (Carhart, 1997; Fama & French, 1993) four-factor exposures[11]. To maintain a robust sample size, factor loadings were estimated over a period of 24 months prior to January 2016 (instead of 60 months as suggested by Lins et al. (2017)). Firms for which data was only available for fewer than 12 months to estimate the factor loadings were omit-

[10] http://business-school.exeter.ac.uk/research/centres/xfi/famafrench/files/

[11] The factor returns were again obtained from the Xfi Centre for Finance & Investment. Stata commands for estimating the factor loadings are provided in the appendix.

ted[12]. This leads to the following regression setup. All control variables are defined in Table 1. We measure financial health indicators and firm size at the end of 2015.

$$\{\text{cross-sectional regr.}\}: \text{BHAR}_i = b_0 + b_1 * \text{CSR}_{i,2014} + b_2' * X_{i,2015} + \text{ indust } + \varepsilon_{it} \quad (2)$$

α	constant
b_1	estimated effect of CSR on shop-period BHARs
$\text{CSR}_{(i,2014)}$	proxy for CSR (linear ESG, ES, or G score) measured at year-end 2014
b_2'	vector of coefficients for the control variables
X_i	vector of control variables lagged by one year
indust	industry fixed effects
ε_{it}	error term

Building on findings of the effect of CSR varying across industries (e.g. Jo & Na, 2012), we include industry dummies at the two-digit ICB industry code level. This further pays credit to findings by Hill et al. (2019); Ramiah et al. (2017); Tielmann and Schiereck (2017) suggesting some industries were more severely affected by the Referendum than others. Further, heteroskedasticity-consistent standard errors were applied.

3.4.2. Corporate Financial Performance Surrounding the Brexit Referendum

The DiD methodology is a shock-based research design for causal inferences aiming at limiting selection and omitted variable bias (Atanasov & Black, 2016, p. 216). A standard DiD design assumes a shock separates firms into treatment and control groups. Thereby, the treatment firms are the ones affected by the shock, whereas the control firms remain completely unaffected (Chava & Purnanandam, 2010, p. 273). This allows to empirically estimate the shock's effects on the treated; the average treatment effects for the treated firms (ATT). They are estimated "as (after-minus-before change for treated firms) minus (after-minus-before change for control group)" (Atanasov & Black, 2016, p. 218):

$$\{\text{panel DiD}\} : y_{it} = \alpha + f_i + g_t + (\delta_{\text{DiD}} * w_{it}) + \varepsilon_{it} \quad (3)$$

α	constant
f_i	firm fixed effects
g_t	time fixed effects
ε_{it}	error term
δ_{DiD}	empirical estimate of ATT
w_{it}	interaction term between a treatment dummy variable (1 for firms in the treatment group; 0 for firms in the control group) and a time dummy (1 if observation is in the event window; 0 if outside of the event window)[13]

In the present case, all UK firms were affected by the Brexit Referendum, independent of their social capital previously accumulated through CSR. Rather, it is attempted to capture a differing sensitivity to policy uncertainty of firms based on their CSR performance. According to Lins et al. (2017) and Chava and Purnanandam (2010), a more related research design is the DiD model with continuous treatment (DiD-continuous), which will be applied in the following analysis. DiD-continuous allows capturing the effect of firms with differing shock sensitivity. It is posited that policy uncertainty affects firms' performance by increasing uncertainty among market participants about companies' future cash flows and financial health. Trust build-up through CSR activities and the concept of reciprocity are hypothesised to alleviate these uncertainties among shareholders, lenders, and customers. Consequently, it is expected that an increase in social capital reduces exposure to policy-induced uncertainty. Thereby, ESG scores as of 2014 provided by ASSET4 are instrumentalised to impose a parametric form on the shock sensitivity[14]. The latter leads to replacing the interaction term w_{it} from eqn. (3) with an interaction between the shock-representing time dummy and the sensitivity-to-shock measure (Atanasov & Black, 2016, p. 253)[15]:

$$\{\text{panel DiD - continuous}\} : y_{it} = \alpha + f_i + g_t + \left(\delta_{\text{DiD}} * \text{ event}_{it} * \text{ sensitivity}_i\right) + \varepsilon_{it} \quad (4)$$

The model setup applied in this study follows Lins et al. (2017) in capturing effects of CSR on firm performance during the shock period, i.e. in 2016, and after the shock period, i.e. from 2017 to 2018. This setup allows us to scrutinise the effects of CSR on financial firm performance in times of increased and moderate policy uncertainty. In this sense, the research design allows detecting whether effects of CSR are unique to times of increased policy uncertainty. Figure 3 illustrates how the sub-sample periods were set out. The following DiD-continuous model was set up:

[12] This was the case for 16 firms.

[13] Atanasov and Black (2016, p. 218) explain in this case the interacted variables do not need to be included separately to estimate the interaction term's coefficient due to firm and time fixed effects.

[14] The theoretical considerations behind the channel through which uncertainty affects firms and continuous variable to measure sensitivity are perceived to be in line with Atanasov and Black (2016, p. 253) "only-through-condition" for DiD-continuous designs.

[15] Atanasov and Black (2016, p. 253) explain a common variation of DiD with a continuous sensitivity measure is to estimate a standard DiD regression using high- and low-sensitivity subsamples indicated with a treatment dummy variable. The treatment dummy equals 1 for high and 0 for low sensitivity (e.g. top vs bottom quarter or third; medium sensitivity is dropped). This variation was not chosen due to concerns of small sample size and selection bias.

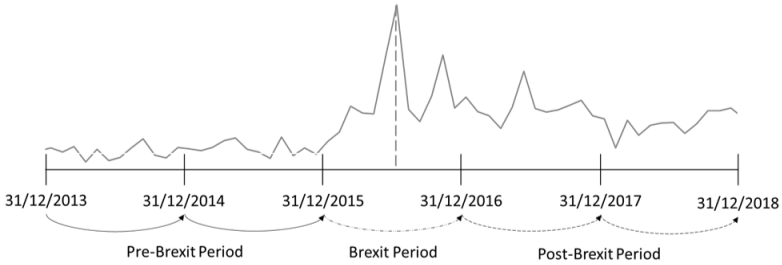


Figure 3: Construction of Sample Periods

$$y_{it} = b_0 + b_1 * \text{CSR}_{i,2014} * \text{Brexit} + b_2 * \text{CSR}_{i,2014} * \text{PostBrexit} + b_3' * X_{i,t-1} + f_i + g_t + \varepsilon_{it} \quad (5)$$

b_0	constant
b_1	Brexit-period interaction term
$\text{CSR}_{(i,2014)}$	proxy for CSR measured at year-end 2014
Brexit	dummy variable set to one in 2016; otherwise zero
b_2	Post-Brexit-period interaction term
PostBrexit	dummy variable set to one in 2017 and 2018; otherwise zero
b_3'	vector of coefficients for the control variables
$X_{(i,t-1)}$	vector of control variables lagged by one year
f_i	firm fixed effects
g_t	time fixed effects
ε_{it}	error term

Firm fixed effects were chosen as an econometric specification to control for unobservable characteristics influencing a particular firm's financial performance (firm fixed effects). Time-fixed effects are included to control for patterns in the performance of all firms, which are likely to persist in times of policy uncertainty (see, inter alia, Davies & Studnicka, 2018; Hill et al., 2019). Both settings aim at reducing the omitted variable bias. Additionally, standard errors are clustered at the firm level. Thereby, a common time-series characteristic also prevalent in panel data is controlled for: the possible temporal serial correlation in error terms by each firm (Petersen, 2009; Thompson, 2011).

The control variables used in the models mainly correspond to the ones used by Buchanan et al. (2018). Additionally, we follow Ding et al. (2016) and also control for stock turnover in the models explaining Tobin's Q and ROE. Table 1 provides an overview.

4. Empirical Results

4.1. Descriptive Statistics

In the appendix, panel A of Table 5 and Table 6 respectively provide the descriptive statistics for all variables. Our primary variables of interest are ESG, ES and Governance scores. From Table 5, we can see that the overall ESG score in 2014, the baseline year of analysis, is rather high with a mean of 71.6, a median of 80, and a cut-off point for the bottom quartile of 55.68. This might indicate UK firms tend to perceive CSR activities as a worthwhile investment, indicating first support for the stakeholder view. This is further enforced by the maximum value being close to 100 across all ESG dimensions. Turning to individual dimensions, Governance exibits the highest values on average, with environmental scores being the lowest. Given that Servaes and Tamayo (2013) argue governance to mainly address shareholder issues, the higher Governance scores could be interpreted to indicate a tilt in CSR activities towards the shareholder view. Turning to Table 5 reveals ESG scores were not persistent during the first half of the sample period from 2014-2016. Overall, ESG scores increased by 16% on average, with ES scores even by 17%. In contrast, governance scores remained relatively stable on average (0.5% increase).

Our first dependent variable to be analysed, BHARs, are on average slightly negative, with the median (-9%) and bottom quartile returns (-15.8%) being strongly negative. Negative BHARs are in line with Davies and Studnicka (2018) finding significant negative abnormal returns surrounding the Brexit Referendum. Together with the fact that the average Z-score during the sample period is below 1.8[16], BHARs indicate stakeholders might have been concerned about some firms' survival prospects. In this sense, it is not surprising that mean and median for Tobin's Q is below 1, suggesting firms are on average undervalued during the sample period (Hirschey, 1985). The latter notion is further supported by the high gross margin and reasonable sales growth exhibited on average.

Panel B of Table 5 and Table 6 show the pairwise correlation coefficients. It is evident that for both cases, correlations between independent variables to be combined in a regression are below 0.65 so that no issues of multicollinearity arise.

In terms of the correlation between the ESG pillars, it has to be noted that there is a high (0.74), statistically significant correlation between the environment and social score. Hence, it can be assumed that regressions with just the environment or just the social score included in the interaction terms will most likely capture similar effects[17]. Thus, CSR is

[16] According to Altman (1993), a score below 1.8 indicates a firm is headed for bankruptcy. However, it has to be noted this cut-off point was estimated in 1993.

[17] This anticipation was corroborated in untabulated estimations of the

split into just two components: a composite environmental and social (ES), and an individual governance (G) score. The two-way separation is further justified by the moderate correlation between G and both the ES score and its individual components.

In a standard DiD design with clearly separated treatment and control groups, it is further needed to defend that the dependent variables in question follow a similar trend for both company groups prior to the event. Thereby, the possibility of the outcome being driven by other unobserved factors is ruled out (Atanasov & Black, 2016). Even though Atanasov, Black, Ciccotello, and Gyoshev (2010) and Lins et al. (2017) do not defend this underlying assumption of DiD in their continuous treatment setups, we still want to ensure a common pre-treatment development prior to the Referendum among firms with high and low sensitivity to the shock. Following our hypotheses, firms with higher ESG scores are theorised to be less sensitive. Analogous to Lambert, Noth, and Schüwer (2017) and L. Nguyen and Wilson (2018), the trend assumption is verified employing a graphical inspection of the dependent variables' mean values for firms with different shock-sensitivity over time. To do so, firms are clustered into tertiles[18] based on their ESG, ES and Governance scores in 2014. Then, the bottom (theorised highest sensitivity) is compared to the top (theorised lowest sensitivity) tertile. Table 7 in the appendix reports the graphs, the majority of which support a shared pre-event trend[19], further supporting eligibility of the DiD methodology.

4.2. Regression Outcomes

4.2.1. Cross-Sectional Regression

Table 2 presents regression estimates of shock-period buy-and-hold-returns (March to September 2016) on linear ESG, ES, and Governance scores as well as different sets of control variables. In models (1) to (3), only stock-market-focused factor loadings are included; financial health and size measures are additionally included in models (4) to (6).

Model (1) exhibits the composite CSR measure is marginally statistically significant at 10% level and negative, providing indicative evidence for a negative effect of previously accrued social capital on the change in shock-period returns. The specification of column (1), however, does not pay credit to effects of firm-specific characteristics that were found in literature to impact firm performance during policy uncertainty. Controlling for this (column (4)), overall CSR's negative effect is more pronounced. Statistical significance increases to the 5% level. Economic significance rises by ca. 1.6%-points[20] given the standard deviation of 2014 ESG scores of 22.962. Hence, a one-standard-deviation increase in 2014 ESG scores causes a reduction in shock-period BHARs by 5.1%-points. Given the mean of shock-period returns of -2.1%, this effect is economically large. Turning to the controls, regression results do not support our anticipation of firms with better financial health having superior stock performance as all financial health variables are insignificant. This is also the case for the size measure. This finding is consistent across all models.

As for the stock-market-centered controls, CSR has a slightly higher economic significance as market risk (4.9%-points) under setup (4). The momentum exposure still remains the most significant influencing factor as a one-standard-deviation-increase leads to a reduction of BHARs by 7.3%-points.

In the following, we disentangle CSR to see if the ES and Governance components are evaluated differently by investing stakeholders. As for the ES pillar, respective coefficients are again negative, but insignificant both without (2) and with (5) controls for financial health and size. In contrast, Governance scores are significant at the 5% level in both setups ((3) and (6)), again with negative coefficients. The impact of governance is also largely economically significant with an impact of -3.54%-points and -4.11%-points respectively per one-standard-deviation-increase (14.173) in pre-Referendum Governance scores. This is further supported by Governance having an impact on returns of more than half of the market risk exposure (5.92%-points) without and almost the same impact (4.35%-points) with firm characteristic and financial health controls. Compared to the ESG factor, Governance has a higher economic impact for a one-standard-deviation-increase in the first model set-up, but turns out to be of lower significance in the second setup.

Taken together, the statistical significance and economic magnitude of ESG and Governance scores on shock-period buy-and-hold-returns indicate accrued social capital indeed offers important explanation for shock-period returns. Given the insignificance of coefficients for ES scores, we infer the effect of ESG is mainly driven by the Governance pillar. The signs of ESG and Governance coefficients signal investors were more concerned about firm performance when pre-Referendum social capital reservoirs were high. One explanation for this might be Burkart, Gromb, and Panunzi (1997) argument of good corporate governance restricting managerial initiative. Finkelstein and D'Aveni (1994), for instance, support this with their finding of powerful CEOs being able to make quicker decisions in uncertain environments. Hence, investors might attribute Good Corporate Governance practices to a reduced ability to withstand policy uncertainty due to managerial restrictions. This is empirically supported by

DiD-models in Table 3 with individual environmental and social scores. They exhibited mostly identical coefficients for the interaction terms.

[18] Contrary to Lambert et al. (2017), we divide firms into tertiles instead of quartiles due to the significantly lower sample size.

[19] For clustering by ESG, the pre-Brexit trend seems to be supported for all dependent variables but ROE and ROA. The same goes for clustering by ES. For Governance, it is Tobin's Q, gross margin, financial slack, and ROA which seem less convincing. We nevertheless perform DiD regression analysis for these dependent variables.

[20] We follow Bouslah et al. (2018, n. 31) and measure economic significance by multiplying the standard deviation of ESG scores as of 2014 by the estimated coefficient associated with the linear ESG measure. For instance, we calculated for ESG: $22.962 * (-0.0022) = -0.051$.

Table 2: Shock-Period Returns and Continuous ESG Scores

Heteroskedasticity-consistent standard errors are presented in parentheses. ***, **, * indicate the estimated coefficient is significantly different from zero at the 0.01, 0.05, and 0.1 level respectively.

VARIABLES	(1) bharw	(2) bharw	(3) bharw	(4) bharw	(5) bharw	(6) bharw
esg_14	-0.0015*			-0.0022**		
	(-1.77)			(-2.03)		
es_14		-0.0012			-0.0014	
		(-1.43)			(-1.37)	
gov_14			-0.0025**			-0.0029**
			(-2.19)			(-2.14)
profw				-0.2541	-0.3202	-0.3112
				(-0.97)	(-1.17)	(-1.14)
ltd_atw				0.1516	0.1719	0.1671
				-1.22	-1.38	-1.35
dlc_atw				-0.07	-0.0854	-0.0586
				(-0.43)	(-0.52)	(-0.35)
cash_taw				0.204	0.2369	0.2553
				-0.64	-0.73	-0.77
beta_rmrfw	0.0817***	0.0794***	0.0755***	0.0625*	0.0614*	0.0555*
	-2.75	-2.68	-2.69	-1.82	-1.78	-1.66
beta_umdw	-0.1025***	-0.1025***	-0.1057***	-0.0860**	-0.0830**	-0.0858**
	(-2.98)	(-2.99)	(-3.17)	(-2.20)	(-2.11)	(-2.26)
beta_hmlw	0.0732***	0.0736***	0.0780***	0.0602***	0.0603***	0.0648***
	-4.03	-4.06	-4.28	-3.02	-3.02	-3.28
beta_smbw	-0.0544**	-0.0524*	-0.0470*	-0.0432	-0.0446	-0.041
	(-2.00)	(-1.92)	(-1.83)	(-1.37)	(-1.41)	(-1.31)
Constant	0.0861	0.0568	0.1918*	-0.0143	0.0451	0.2217
	-1.06	-0.72	-1.85	(-0.06)	-0.19	-1.02
Industry dummies	Yes	Yes	Yes	Yes	Yes	Yes
Observations	308	308	308	267	267	267
R-squared	0.331	0.326	0.332	0.361	0.354	0.362

van van Essen, Engelen, and Carney (2013) showing Governance was negatively related to firm performance during the 2008 financial crisis. Lins et al. (2017) corroborate the same effect on stock returns.

Findings indicate investors required a risk premium for CSR investments aggregated across CSR components and for Good Governance additionally to a policy uncertainty risk premium (Belo et al., 2013; Erb et al., 1996). In this sense, our findings suggest the rejection of H_0^1 of social capital being positively related to shock-period stock returns while providing support in favour of H_0^5.

4.2.2. Difference-in-Difference Regressions

We next scrutinise whether non-investing stakeholders hold positive beliefs about firms investing in CSR, which translate into increased real performance on the level of operating performance, financial health, and firm value surrounding the Brexit Referendum. In the models presented in Table 3, the variables of interest are the two interaction terms between CSR measures and time dummies. The coefficient of the interaction term "ibr" captures the differential impact of the respective CSR measure on the scrutinised corporate financial performance measure in the year of the Referendum after controlling for firm characteristics and after removing the firms' average performance (firm fixed effects) over the entire estimation period and after removing time-series patterns (time-fixed effects) among the overall financial performance measure (Lins et al., 2017). In this sense, the coefficient of "ibr" measures marginal effects of CSR in times of increased economic policy uncertainty. The coefficient of "ipost" captures the equivalent effect for times of moderated policy uncertainty.

If effects of CSR are more pronounced during increased policy uncertainty, we should observe a significant coefficient for "ibr" interaction terms as well as a reversal of the effect in the post-Referendum period (Buchanan et al., 2018).

Panel A of Table 3 presents regression results regarding the composite ESG measure. Looking at operating performance, we first analyse distribution as an influencing channel for CSR. We thereby investigate whether customers were

more willing to stick with a company due to previously-build, positive beliefs about them. Here, our first model in column (1) scrutinises whether higher-CSR firms exhibit higher changes in gross margins surrounding the Referendum. Neither interaction term shows a statistically significant coefficient at 5%, indicating higher-CSR firms did not realise higher mark-ups around the Referendum. We next turn to sales growth (2) and again find no significant effect at 5% in both time periods. Cognizant that there are no differences between high and low-CSR performers in 2016 and post-Referendum, it is not surprising that we cannot detect any significant effect of CSR on profitability during high and moderated policy uncertainty.[21] This finding holds for both ROA (column (3)) and ROE (column (4)). Hence we can infer high-CSR firms neither exhibit increased operating performance relative to other firms during high nor during moderated policy uncertainty. Our findings are in line with Nelling and Webb (2009), but contradict Guenster et al. (2011).

Our next set of models focuses on financial health. The first explained variable is financial slack (5), i.e. the cash available to finance new projects. We find overall CSR performance to have no significant effect on the cash cusion of firms to withstand tough times. The model on Altman's Z (6) indicates ESG neither has a stabilising nor destabilising effect on firms during or after 2016. Given the two variables are in part derived from explained variables in the models on operating performance, the findings are not surprising. In sum, we cannot support Goss and Roberts (2011) finding of reduced default risk.

Column (7) exhibits our last level of real performance: firm value. Tobin's Q further offers a forward-looking market perspective. Additionally, it captures the market value managers can generate per unit of underlying asset so that Tobin's Q also assesses managerial quality (Guenster et al., 2011). Nevertheless, the market perspective is also influenced by operating performance and financial health criteria. In this sense, findings of insignificant interaction term coefficients for both subsample periods are in line with the previous model settings. In this finding, we contradict Ghoul et al. (2017); Guenster et al. (2011); Jiao (2010); Nelling and Webb (2009).

Among the control variables, it might be worth noting that size exhibits highly significant negative effects at 1% level on sales growth and operating profitability measured by ROA, as well as financial slack over the whole sample period. Further, firms with higher operating performance measured by ROA and sales growth seem to exhibit higher financial slack (5) and financial stability (6). Higher sales growth is also significantly positively linked to higher gross margins (1). These findings, however, apply for the whole sample period and are but indicative findings for effects during policy uncertainty. Nevertheless, they seem to generally be in line with Davies and Studnicka (2018) and Hill et al. (2019) investigations on Brexit exposure and, more generally, on impacts of policy uncertainty (inter alia, Boutchkova et al., 2012; Gulen & Ion, 2016).

To address hypothesis number five, we reestimate the DiD-models but interact the Brexit and PostBrexit dummies with aggregated ES (Table 3 Panel B) and Governance scores (Table 3 Panel C) respectively. Regarding the main variables of interest, insignificant effects across all models and both interaction terms remain persistent. Further, significance and sign of aforementioned controls are consistent across all panels. Note that the correlation between ES and Governance scores as of 2014 is significantly below 0.65 so that consistent results were not necessarily anticipated. Hence, we cannot corroborate our previous finding of Governance having more significant impacts on financial performance than ES. In contrast to investors, non-investing stakeholders of UK-listed firms did not exhibit the selective value attribution that was found by Harjoto and Jo (2015); Jiao (2010); Kim et al. (2014); Nelling and Webb (2009) for the US. In this sense, no support is provided for H_0^5 on the level of real performance; however, we find evidence against it on the level of stock performance.

In summary, the DiD models reveal two key findings. First, the benefits of relational wealth previously created through the sum of CSR as well as on the levels of ES and Governance do not outweigh their costs during increased policy uncertainty in 2016. In context with our theoretical framework, we conclude neither the effects of the reputation-building benefits nor overinvestment costs prevailed. More specifically, relational wealth along distribution channels did not have significant effects. At the same time, managers of high-CSR firms are not indicated to have exploited incomplete contracts for intrinsic interests to an extent harming real performance sourrounding the Referendum. In sum, high-CSR firms' real performance is neither less nor more sensitive to Brexit-induced policy uncertainty.

The second finding is the effect consistency across the shock and post-shock periods. Hence, we were not able to detect a dynamic relationship between CSR varying across times of increased and moderated economic policy uncertainty. Hence, our model outputs contradict what was initially anticipated. However, this finding might be attributable to the fact that policy uncertainty levels after the Referendum are slightly lower, but still mostly above the overall and ten-years mean. The reason might be that Brexit procedures are still ongoing without any agreements on a Brexit deal between the UK and the EU. In fact, commentators perceive a no-deal Brexit which would leave the UK without access to the EU free-trade zone to be ever more likely (Wilson, 2019). Hence, effects of policy uncertainty are still well persistent in 2017 and 2018 (Figure 1). With that said, the magnitude of the reduction in policy uncertainty might not be strong enough. Alternatively, changes in policy uncertainty might exhibit a response lag effect in significantly changing stake-

[21] We further investigate effects of ESG, ES, and Governance on the customer channel in untabulated regressions leading to consistent fndings. We regress accounts receivables over sales against the same explanatory variables but do not find customers of high-CSR frims to pay their accounts faster in the year of the Referendum or after. Further, we investigate whether employees are more productive by regressing sales per employee. We again fail to detect any significant relationship.

Table 3: Real Performance and CSR Surrounding the Brexit Referendum

Standard errors clustered at firm level are reported in parentheses. ***, **, * indicate the estimated coefficient is significantly different from zero at the 0.01, 0.05, and 0.1 level respectively.

Panel A: Effects of Composite ESG Scores on Real Performance in the year of and after the Brexit Referendum

VARIABLES	(1) gmw	(2) sales_gw	(3) roa_netincw	(4) roew	(5) s_cashw	(6) zscorew	(7) tq2w
ibresg	0.0001	-0.0003	0.0000	0.0000	-0.0001	-0.0001	0.0000
	(0.30)	(-0.48)	(-0.06)	(-0.62)	(-0.67)	(-0.05)	(0.08)
ipostesg	-0.0001	-0.0007	-0.0001	-0.0001	0.0000	0.0012	-0.0003
	(-0.42)	(-1.07)	(-0.22)	(-1.31)	(0.02)	(1.06)	(-0.94)
lagsizew	-0.0272*	-0.0978***	-0.0685***	-0.0006	-0.0453***	-0.1257*	-0.0005
	(-1.77)	(-2.91)	(-4.82)	(-0.33)	(-5.23)	(-1.69)	(-0.03)
lagcapexpw	0.0637	-1.6932***	-0.2582**	-0.0175	-0.0956	-1.7353***	0.3153**
	(0.48)	(-5.67)	(-2.01)	(-1.02)	(-1.05)	(-2.73)	(2.07)
lagfabaw	0.0390	0.1420	0.0068	0.0036	-0.0723*	0.5574	0.0320
	(0.49)	(0.84)	(0.08)	(0.19)	(-1.68)	(1.64)	(0.36)
lagcash_taw	0.0515	-0.1543	0.0987*	0.0104	-0.0044	0.2922	0.0357
	(0.87)	(-0.94)	(1.66)	(1.29)	(-0.10)	(0.72)	(0.43)
lagbleveragew	-0.0414	0.0810	0.1147**	0.0028	-0.0471	-0.2071	0.3032***
	(-0.75)	(0.82)	(2.45)	(0.26)	(-1.64)	(-1.12)	(5.60)
lagrdw	-0.7948	-0.4914	-0.0055	0.0304	-0.0902	0.3270	-0.2107
	(-1.42)	(-0.58)	(-0.02)	(1.49)	(-0.34)	(0.22)	(-0.63)
lagadvertisingw	0.0813	0.3491***	0.0021	-0.0083	-0.0232	-0.1158	-0.0089
	(0.74)	(2.85)	(0.03)	(-1.11)	(-0.68)	(-0.47)	(-0.15)
turnoverw				-0.0030*			-0.0033
				(-1.75)			(-0.28)
lagroa_netincw	0.0654	0.0819		-0.0021	0.0488**	0.5627***	-0.0810
	(1.38)	(0.81)		(-0.38)	(2.22)	(2.92)	(-1.44)
lagsales_gw	0.0727***		0.0441*	0.0055	0.0327**	0.2720***	0.0013
	(3.06)		(1.67)	(1.18)	(2.44)	(3.25)	(0.05)
Constant	0.8079***	1.3767***	0.9935***	-0.0018	0.7780***	3.5795***	0.4231
	(3.57)	(2.66)	(4.69)	(-0.06)	(6.01)	(3.22)	(1.44)
Firm fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,310	1,313	1,309	1,256	1,308	1,278	1,255
R-squared	0.046	0.16	0.097	0.035	0.09	0.126	0.115
Firms included	283	284	283	274	283	278	274

Panel B: Effects of Composite ES Scores on Real Performance in the year of and after the Brexit Referendum

VARIABLES	(1) gmw	(2) sales_gw	(3) roa_netincw	(4) roew	(5) s_cashw	(6) zscorew	(7) tq2w
ibres	0.0001	-0.0004	0.0000	0.0000	0.0000	0.0006	0.0001
	(0.40)	(-0.68)	(0.15)	(-0.54)	(0.09)	(0.51)	(0.34)
ipostes	0.0000	-0.0004	0.0001	-0.0001	0.0001	0.0020*	-0.0002
	(-0.05)	(-0.64)	(0.34)	(-1.18)	(0.93)	(1.67)	(-0.75)
lagsizew	-0.0271*	-0.0983***	-0.0680***	-0.0008	-0.0448***	-0.1206	-0.0012
	(-1.74)	(-2.89)	(-4.73)	(-0.39)	(-5.22)	(-1.65)	(-0.06)
lagcapexpw	0.0645	-1.6902***	-0.2581**	-0.0172	-0.0981	-1.7558***	0.3159**
	(0.49)	(-5.63)	(-1.99)	(-1.01)	(-1.08)	(-2.80)	(2.08)

(Continued)

Table 3—continued

lagfabaw	0.0423	0.1513	0.0115	0.0045	-0.0666	0.5822*	0.0346
	(0.52)	(0.89)	(0.14)	(0.24)	(-1.55)	(1.73)	(0.38)
lagcash_taw	0.0493	-0.1599	0.0955	0.0097	-0.0072	0.2640	0.0348
	(0.83)	(-0.96)	(1.62)	(1.27)	(-0.16)	(0.65)	(0.42)
lagbleveragew	-0.0423	0.0785	0.1134**	0.0025	-0.0487*	-0.2181	0.3030***
	(-0.76)	(0.79)	(2.41)	(0.23)	(-1.67)	(-1.18)	(5.61)
lagrdw	-0.8118	-0.5133	-0.0183	0.0274	-0.0938	0.2748	-0.2217
	(-1.43)	(-0.60)	(-0.07)	(1.39)	(-0.35)	(0.19)	(-0.67)
lagadvertisingw	0.0831	0.3525***	0.0043	-0.0078	-0.0208	-0.1070	-0.0066
	(0.75)	(2.87)	(0.06)	(-1.06)	(-0.61)	(-0.44)	(-0.11)
turnoverw				-0.0030*			-0.0031
				(-1.73)			(-0.26)
lagroa_netincw	0.0650	0.0800		-0.0021	0.0482**	0.5563***	-0.0795
	(1.35)	(0.79)		(-0.37)	(2.18)	(2.90)	(-1.42)
lagsales_gw	0.0732***		0.0444*	0.0057	0.0327**	0.2704***	0.0017
	(3.02)		(1.67)	(1.22)	(2.46)	(3.24)	(0.07)
Constant	0.8057***	1.3822***	0.9863***	0.0003	0.7692***	3.5070***	0.4322
	(3.53)	(2.64)	(4.62)	(0.01)	(6.00)	(3.19)	(1.46)
Firm fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,310	1,313	1,309	1,256	1,308	1,278	1,255
R-squared	0.045	0.159	0.097	0.032	0.09	0.129	0.115
Firms included	283	284	283	274	283	278	274

Panel C: Effects of Governance Scores on Real Performance in the year of and after the Brexit Referendum

VARIABLES	(1) gmw	(2) sales_gw	(3) roa_netincw	(4) roew	(5) s_cashw	(6) zscorew	(7) tq2w
ibrgov	0.0000	-0.0000	0.0000	-0.0001	-0.0003	0.0006	-0.0002
	(0.04)	(-0.06)	(-0.00)	(-1.21)	(-1.09)	(0.51)	(-0.55)
ipostgov	0.0001	-0.0011	0.0001	-0.0001	-0.0001	0.0013	-0.0005
	(0.30)	(-1.41)	(0.26)	(-0.99)	(-0.54)	(0.86)	(-1.10)
lagsizew	-0.0266*	-0.1015***	-0.0680***	-0.0009	-0.0459***	-0.1209	-0.0022
	(-1.72)	(-2.98)	(-4.61)	(-0.42)	(-5.28)	(-1.64)	(-0.12)
lagcapexpw	0.0689	-1.7314***	-0.2545**	-0.0193	-0.1000	-1.7010***	0.3013**
	(0.51)	(-5.61)	(-1.99)	(-1.08)	(-1.09)	(-2.64)	(1.97)
lagfabaw	0.0444	0.1411	0.0106	0.0049	-0.0733*	0.5501	0.0293
	(0.57)	(0.82)	(0.12)	(0.26)	(-1.76)	(1.61)	(0.33)
lagcash_taw	0.0478	-0.1603	0.0969	0.0088	-0.0032	0.3141	0.0338
	(0.79)	(-0.98)	(1.65)	(1.18)	(-0.07)	(0.77)	(0.40)
lagbleveragew	-0.0428	0.0772	0.1142**	0.0023	-0.0461	-0.1994	0.3027***
	(-0.77)	(0.77)	(2.43)	(0.21)	(-1.60)	(-1.09)	(5.58)
lagrdw	-0.8176	-0.5071	-0.0142	0.0206	-0.0915	0.4681	-0.2370
	(-1.45)	(-0.59)	(-0.05)	(1.16)	(-0.36)	(0.32)	(-0.72)
lagadvertisingw	0.0826	0.3518***	0.0036	-0.0078	-0.0239	-0.1197	-0.0101
	(0.75)	(2.77)	(0.05)	(-1.06)	(-0.71)	(-0.48)	(-0.18)
turnoverw				-0.0030*			-0.0036
				(-1.75)			(-0.30)
lagroa_netincw	0.0632	0.0855		-0.0021	0.0502**	0.5631***	-0.0803
	(1.33)	(0.85)		(-0.38)	(2.27)	(2.91)	(-1.43)

(Continued)

Table 3—continued

lagsales_gw	0.0738***		0.0442*	0.0059	0.0327**	0.2641***	0.0034
	(3.03)		(1.65)	(1.24)	(2.48)	(3.14)	(0.14)
Constant	0.7994***	1.4342***	0.9852***	0.0015	0.7864***	3.5028***	0.4491
	(3.50)	(2.72)	(4.48)	(0.05)	(6.05)	(3.18)	(1.51)
Firm fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,310	1,313	1,309	1,256	1,308	1,278	1,255
R-squared	0.045	0.161	0.097	0.031	0.09	0.124	0.115
Firms included	283	284	283	274	283	278	274

holder behaviour in a way that affects the relationship between ESG and corporate financial performance.

Further discussion in light of previous shock-based studies will be provided in section 6. In sum, both key findings contradict hypotheses H_0^2 to H_0^4.

5. Robustness Tests

5.1. Cross-Sectional Regression

Following Lins et al. (2017), we reestimate our baseline regressions in Table 8 in the appendix and replace the continuous CSR measures by dummies indicating the top (t3) and medium tertile (t2). The bottom tertile is captured by the constant. We thereby evaluate whether our previous finding of inverse effects on BHARs are more pronounced at higher levels of ESG and Governance scores.

The robustness test reveals negative effects of higher levels of pre-event social capital are robust for the Governance pillar, but not for the composite ES and ESG levels. This indicates effects of ES and ESG depend upon econometric specifications and are thus less pronounced as indicated by the baseline model.

We also conduct a further robustness test put forth by Lins et al. (2017) and measure CSR at different points in time to see whether our results are driven by the point in time of measuring CSR. Alternatively, we measure CSR in 2015 as another pre-shock measure, and in 2016. Results are reported in Table 9 and Table 10 in the appendix. In 2015, all variables of interest are not significant at 0.05 level; in 2016, the ESG coefficient is only negative and significant with stock-market-based controls, while ES is negative and significant in both model setups. The findings contradicting the baseline model are not surprising given the high increases in ESG, ES, and Governance scores between 2014 and 2016, mainly for low-tier firms. We explain this with Godfrey (2005) theory of firms instrumentalising CSR engagement as a risk management tool in anticipation of turbulent times (Buchanan et al., 2018; Godfrey, 2005; Godfrey et al., 2009; Lins et al., 2017). This is indicated in the descriptive statistics by low-tier cutoff-points for ESG quadrupling. For Govenance, the increase was even sevenfold. As low-tier firms moved closer to the average, kurtosis and negative skewness increased, further supporting this indication. Further research is needed, though, to establish a causality. Under these considerations, insignificance of Governance scores might hint at increased Good Governance efforts in times of increasing policy uncertainty are appreciated by investors as a means of reducing agency conflicts (Cui et al., 2018; Eccles et al., 2014; Knack & Keefer, 1997; Putnam, 1993), possibly to mitigate increased cash flow uncertainty (Cremers & Yan, 2016). In contrast, overinvestment concerns arise among investors for increased combined environmental and social endeavours (Surroca & Tribó, 2008). This indication, again, requires further research.

5.2. Difference-in-Difference Models

Firstly, we again reestimate basline DiD-models using CSR measures as of 2015 and 2016. The regression results are reported in the appendix. In summary, the variables of interest, the two interaction terms, remain insignificant at the 5% level. Consequently, the results mainly reinforce the previous findings of ESG, ES, and Governance having no significant impact on real performance on the level of operating profitability, financial health, and overall firm value.

In a second set of untabulated robustness tests, we test for alternative variable definitions. As for dependent variables, we reestimate baseline DiD-models for Tobin's Q and ROA using the following alternative definitions:

$$\text{TQ} = \frac{\text{market value of equity}_t + \text{ total liabilities}_t}{\text{book value of equity}_t + \text{ total liabilities}_t} \quad (6)$$

$$\text{ROA}_{\text{ebit}} = \frac{\text{ebit}_t}{\text{total assets}_t} \quad (7)$$

Our main findings remain unchanged. This also holds when reestimating all 21 baseline regressions using the market instead of the book leverage. Market leverage was calculated as follows:

$$\text{ML} = \frac{\text{long-term debt}_t + \text{current liabilities}_t}{\text{total asset-common equity}_t + \text{market capitalisation}_t} \quad (8)$$

The same goes for reestimating all 21 baseline regressions with the natural logarithm of market capitalisation and

revenues, respectively as alternative size measures. We can summarise the findings from the DiD-models are robust to alternative variable definitions.

6. Further Discussion on the Dynamic Relationship Between CSR and Corporate Financial Performance

This paper contributes to literature on the dynamic relationship between CSR and corporate financial performance with three main findings: (1) CSR and especially Governance are inversely related to shock-period returns, (2) high-CSR and low-CSR firms' real performance does not differ in terms of sensitivity to adverse Brexit shock implications, and (3) effects of CSR on real performance do not vary alongside levels of policy uncertainty. We now proceed to discuss our findings in context with the three papers most closely related to our work.

These three findings are in contrast to Bouslah et al. (2018); Buchanan et al. (2018); Lins et al. (2017), all finding effects of CSR to be more pronounced during an adverse shock to markets, the 2008 financial crisis. Putting our findings in context with these papers, we identify two mechanisms that could contribute to the disparity.

This paper resumes as a first inference that previous findings do not necessarily hold in other markets and / or for aggregated shock events other than financial market shocks. Buchanan et al. (2018, p. 82) already hinted at this notion. They found firms engaging in CSR to not exhibit different firm values on average compared to non-CSR firms immediately after the financial crisis. They trace this back to increases in policy uncertainty, given the substantial quantitative easing rounds used by the US Federal Reserve to stabilise the economy. Similarities between the uncertainty situation in the US after the financial crisis and the Referendum are evident: in both cases, prolongued economic downturns and long-term declines in household consumptions were expected (inter alia, Dhingra et al., 2016; Fraser of Allander Institute, 2016; Sampson et al., 2016). The severity of the Referendum's threat to the economy was further illustrated by the dramatic drop of the British Pound Sterling to US Dollar exchange rate (Allen, Treanor, & Goodley, 2016). It was so severe that the Bank of England reacted with lowering the base interest rate from 0.5% to 0.25% two months after the Referendum to mitigate financial pressure on the real economy (Bank of England, 2016). Aggregated shocks to economic policy uncertainty might be of such a structural macroeconomic nature that effects of CSR on real performance diminish. As for the crosssectional regression, the horizon for returns might have been too short-sighted to capture this effect.

Opposing, Lins et al. (2017, p. 1816) detect positive effects of CSR on returns and operating performance still persist in the years after the crisis from 2010 to 2013. Their finding leaves us to suggest as a second inference that there is another, complemental factor influencing the effect of varying macro conditions on the relationship between CSR and firm performance: aggregated levels of trust among stakeholders towards firms. Aggregated trust levels in the US stayed relatively low after the 2008 financial crisis, which Lins et al. (2017) attribute to be the reason behind the lacking reversal in stock return valuation premia and operating outperformance for high-CSR firms[22]. In contrast, overall trust levels and especially trust in companies in the UK have reached a new hight in 2013, marking the highest level since the 2008 financial crisis. Ever since 2013, trust levels have mostly stagnated according to the Edelman Trust Barometer (Daniel J Edelman Ltd., 2018). Edelman scores for trust in businesses range from 46 to 43 between 2016 to 2018, compared to 36 in 2008. In this sense, our findings are indicative support for Sapienza and Zingales (2012) viewpoint that trust in companies becomes more important when overall levels of trust in markets, institutions, and corporations are low. Hence, presented empirical findings could be seen as further indication for links between financial payoffs to social capital being more pronounced in times of low aggregated levels of trust as established by Guiso et al. (2008) and Lins et al. (2017).

7. Conclusion

This paper examines the impact of economic policy uncertainty on the relationship between corporate financial performance and CSR surrounding the 2016 Brexit Referendum. The investigation focuses on the 320 non-financial firms listed at the London Stock Exchange which are part of the ASSET4 ESG database as of 2014. The sample period covers the years of 2014 to 2018 with 2016 indicating increased, and 2017 to 2018 indicating moderated policy uncertainty. As a measure of previously accrued social capital, ESG, ES, and G scores as of 2014 were employed. The main results can be summarised as follows. First, we find moderate empirical support for social capital previously accrued through CSR initiatives to negatively affect buy-and-hold shock-period returns. The negative impact is driven by the Governance score. Effects are statistically and economically meaningful. Results are, however, not robust to capturing sensitivity by using tertile dummies (for ESG) or using ESG, ES, and G scores as of 2015 and 2016.

Second, we do not find non-investing stakeholders value relational wealth created through CSR to an extent that significantly affects the real performance of firms during the year of the Referendum. This holds across both levels of CSR analysed and across operational performance, financial health, and firm value.

Third, we do not confirm previous findings of a dynamic relationship between CSR and corporate financial performance across times of increased (2016) and of moderated policy uncertainty (2017 and 2018). The last two findings are robust to using ESG, ES, and G scores measured at alternative points in time.

[22] The Financial Trust Index indicates 11% and 12% of respondents trusted the stock market and large corporations respectively in 2009. Responses only changed moderately to 15% and 17% at the end of 2013 (Sapienza, Zingales, & Jones, 2019).

We interpret that from the perspective of investing stakeholders, the shareholder view's overinvestment hypothesis was supported for the Brexit period. In contrast, from the perspective of non-investing shareholder and effects on real performance, neither the shareholder nor the stakeholder argument outweighted the other. Consequently, social capital reservoirs are a firm characteristic that increases stock return's but not real performance's sensitivity to Brexit uncertainty.

In the context of the emerging literature on the time-varying nature of CSR effects, we cannot infer that policy uncertainty caused by Brexit is a macro condition that alters effects of social capital in a way that they are more pronounced than during moderate policy uncertainty. This might be because CSR effects are more sensitive to rather short-lived, market-related shocks, but less sensitive to prolonged financial frictions (Buchanan et al., 2018). This indication could be further tested by repeating the conducted analysis once Brexit-induced uncertainty has terminated, i.e. once the UK has officially withdrawn from the EU and has recovered from immediate macroeconomic effects. A second inference discussed is that macro shocks might only influence the (dis-)equilibrium between reputation-building benefits and overinvestment costs if overall levels of trust in companies decline (Lins et al., 2017).

However, these inferences have to be interpreted with caution as, so far, no other study has analysed impacts of policy uncertainty shocks on the relationship between CSR and corporate financial performance. Therefore, it is suggested to perform further robustness tests in the form of alternative research designs. One example could be the standard DiD setup employed by Buchanan et al. (2018). Thereby, firms with ESG scores are assigned to the treatment and firms without to the control group. This would require to create a matched sample to ensure covariate bias to establish DiD credibility (Atanasov & Black, 2016). A frequently employed technique is nearest-neighbour propensity score matching with common support (e.g. Buchanan et al., 2018; Lambert et al., 2017; L. Nguyen & Wilson, 2018)[23]. When applying the matching technique, it might also be interesting to estimate ATTs using propensity score matching[24] instead of DiD regression estimation. Atanasov and Black (2016) further suggest exploying shock Instrumental Variable and Regression Discontinuity designs as additional robustness tests for DiD. They also mention placebo tests as possible robustness tests but suggest they are more relevant to studies on legal shocks.

Alternatively, ESG measures of different data providers such as Bloomberg or MSCI could be used for robustness checks.

Given that shock-based research on CSR is still emerging, this study provides several indications for further research. In terms of the Brexit event, it is first suggested to apply the DiD-methodology also for stock returns. A second proposal is to implement a triple-differencein-difference setup in which CSR and the time dummies further interact with either firm size, internationalisation, maturity, or performance. Theoretical justification is that Davies and Studnicka (2018) and Hill et al. (2019) find firms differing on these characteristics also differ in their exposure to Brexit uncertainty. Additionally, Chan, Watson, and Woodliff (2014) and Gamerschlag, Möller, and Verbeeten (2011) indicate larger firms have higher ESG activities, whereas Hong et al. (2012) report the same effect for more profitable firms.

Another interesting direction is to look at different types of aggregated shock events, e.g. additional shocks of policy uncertainty such as 9/11, legal shocks such as the withdrawal of the United States form the Paris climate change agreement, or further market shocks such as the dot-com bubble. This would be useful to establish criteria for shocks that are likely to influence the (dis-)equilibrium between shareholder and stakeholder view of the firm. Related to this is the proposition to conduct the analysis on different countries since Ghoul et al. (2017) provide evidence for incremental value of CSR being stronger in countries with reduced market-supporting institutions.

[23] Stata code for performing nearest-neighbour propensity score matching is provided in the appendix.

[24] This is done in Stata using the command *teffects*.

References

Allen, K., Treanor, J., & Goodley, S. (2016, jun). *Pound slumps to 31-year low following Brexit vote*. Retrieved from https://www.theguardian.com/business/2016/jun/23/british-pound-given-boost-by-projected-remain-win-in-eu-referendum

Altman, E. (1993). *Corporate financial distress and bankruptcy*. New York: Wiley.

An, H., Chen, Y., Luo, D., & Zhang, T. (2016). Political uncertainty and corporate investment: Evidence from China. *Journal of Corporate Finance*, *36*, 174–189. doi: 10.1016/j.jcorpfin.2015.11.003M4-Citavi

Ansoff, H. I. (1965). *Corporate Strategy*. New York: McGraw-Hill M4 - Citavi, New York.

Arora, P., & Dharwadkar, R. (2011). Corporate Governance and Corporate Social Responsibility (CSR): The Moderating Roles of Attainment Discrepancy and Organizational Slack. *Corporate Governance: An International Review*, *19*(2), 136–152.

Atanasov, V., & Black, B. (2016). Shock-Based Causal Inference in Corporate Finance and Accounting Research. *Critical Finance Review*, *5*, 207–304.

Atanasov, V., Black, B., Ciccotello, C., & Gyoshev, S. (2010). How does law affect finance? An examination of equity tunneling in Bulgaria. *Journal of Financial Economics*, *96*(1), 155–173. doi: 10.1016/j.jfineco.2009.12.005

Baker, S. R., Bloom, N., & Davis, S. J. (2016, nov). Measuring Economic Policy Uncertainty. *The Quarterly Journal of Economics*, *131*(4), 1593–1636. Retrieved from https://academic.oup.com/qje/article-lookup/doi/10.1093/qje/qjw024 doi: 10.1093/qje/qjw024

Baker, S. R., Bloom, N., & Davis, S. J. (2019). *Economic Policy Uncertainty*. Retrieved 2019-07-12, from https://www.policyuncertainty.com/about.html

Bank of England. (2016). *The Interest Rate (Bank Rate)*. Retrieved from https://www.bankofengland.co.uk/boeapps/database/Bank-Rate.asp

Barnea, A., & Rubin, A. (2010). Corporate Social Responsibility as a Conflict Between Shareholders. *Journal of Business Ethics : JBE*, *97*(1), 71–86. doi: 10.1007/s10551-010-0496-zM4-Citavi

BBC. (2019, feb). *Brexit: Third of members considering move abroad – IOD survey*. Retrieved from https://www.bbc.co.uk/news/business-47083214

Beaulieu, M.-C., Cosset, J.-C., & Essaddam, N. (2005). The impact of political risk on the volatility of stock returns: the Case of Canada. *Journal of International Business Studies*(36), 701–718.

Belo, F., Gala, V. D., & Li, J. (2013). Government spending, political cycles, and the cross section of stock returns. *Journal of Financial Economics*, *107*(2), 305–324. doi: 10.1016/j.jfineco.2012.08.016M4-Citavi

Białkowski, J., Gottschalk, K., & Wisniewski, T. P. (2008). Stock market volatility around national elections. *Journal of Banking & Finance*, *32*(9), 1941–1953. doi: 10.1016/j.jbankfin.2007.12.021M4-Citavi

Bird, R., D. Hall, A., Momentè, F., & Reggiani, F. (2007). What Corporate Social Responsibility Activities are Valued by the Market? *Journal of Business Ethics : JBE*, *76*(2), 189–206. doi: 10.1007/s10551-006-9268-1M4-Citavi

Bloom, N., Bond, S., & van Reenen, J. (2007). Uncertainty and Investment Dynamics. *The Review of Economic Studies*, *74*, 391–415.

Bouslah, K., Kryzanowski, L., & M'Zali, B. (2013). The impact of the dimensions of social performance on firm risk. *Journal of Banking & Finance*, *37*(4), 1258–1273. doi: 10.1016/j.jbankfin.2012.12.004M4-Citavi

Bouslah, K., Kryzanowski, L., & M'Zali, B. (2018). Social Performance and Firm Risk: Impact of the Financial Crisis. *Journal of Business Ethics : JBE*, *149*(3), 643–669. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6449062 doi: 10.1007/s10551-016-3017-xPM-31007319

Boutchkova, M., Doshi, H., Durnev, A., & Molchanov, A. (2012). Precarious Politics and Return Volatility. *Review of Financial Studies*, *25*(4), 1111–1154. doi: 10.1093/rfs/hhr100M4-Citavi

Brammer, S., Brooks, C., & Pavelin, S. (2006, sep). Corporate Social Performance and Stock Returns: UK Evidence from Disaggregate Measures. *Financial Management*, *35*(3), 97–116. Retrieved from http://doi.wiley.com/10.1111/j.1755-053X.2006.tb00149.x doi: 10.1111/j.1755-053X.2006.tb00149.x

Brooks, C. (2014). *Introductory Econometrics for Finance* (3rd ed.). Cambridge University Press.

Buchanan, B., Cao, C. X., & Chen, C. (2018). Corporate social responsibility, firm value, and influential institutional ownership. *Journal of Corporate Finance*, *52*, 73–95. doi: 10.1016/j.jcorpfin.2018.07.004M4-Citavi

Burkart, M., Gromb, D., & Panunzi, F. (1997). Large Shareholders, Monitoring, and the Value of the Firm. *Quarterly Journal of Economics*, 693–728.

Carhart, M. M. (1997, mar). On Persistence in Mutual Fund Performance. *The Journal of Finance*, *52*(1), 57–82. doi: 10.1111/j.1540-6261.1997.tb03808.x

Cespa, G., & Cestone, G. (2007). Corporate Social Responsibility and Managerial Entrenchment. *Journal of Economics & Management Strategy*, *16*(3), 741–771. doi: 10.1111/j.1530-9134.2007.00156.xM4-Citavi

Chan, M. C., Watson, J., & Woodliff, D. (2014, nov). Corporate Governance Quality and CSR Disclosures. *Journal of Business Ethics*, *125*(1), 59–73. Retrieved from http://link.springer.com/10.1007/s10551-013-1887-8 doi: 10.1007/s10551-013-1887-8

Chava, S., & Purnanandam, A. (2010, aug). CEOs versus CFOs: Incentives and corporate policies. *Journal of Financial Economics*, *97*(2), 263–278. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S0304405X10000772 doi: 10.1016/j.jfineco.2010.03.018

Cheng, B., Ioannou, I., & Serafeim, G. (2014). Corporate social responsibility and access to finance. *Strategic Management Journal*, *35*(1), 1–23. doi: 10.1002/smj.2131

Coleman, J. S. (1988). Social Capital in the Creation of Human Capital. *American Journal of Sociology*, *94*(Supplement), 95–120.

Cremers, M., & Yan, H. (2016). Uncertainty and Valuations. *Critical Finance Review*, *5*(1), 85–128. doi: 10.1561/104.00000020M4-Citavi

Crow, D. (2019, jan). *Barclays free to move € 190bn of assets to Ireland ahead of Brexit | Financial Times*. Retrieved from https://www.ft.com/content/dbe4fd26-24ac-11e9-8ce6-5db4543da632

Cui, J., Jo, H., & Na, H. (2018). Does Corporate Social Responsibility Affect Information Asymmetry? *Journal of Business Ethics : JBE*, *148*(3), 549–572. doi: 10.1007/s10551-015-3003-8M4-Citavi

Daniel J Edelman Ltd. (2018). *Edelman Trust Barometer 2018 - UK Findings*.

Davies, R. B., & Studnicka, Z. (2018). The heterogeneous impact of Brexit: Early indications from the FTSE. *European Economic Review*, *110*, 1–17. doi: 10.1016/j.euroecorev.2018.08.003M4-Citavi

Derrien, F., & Kecskés, A. (2013, aug). The Real Effects of Financial Shocks: Evidence from Exogenous Changes in Analyst Coverage. *The Journal of Finance*, *68*(4), 1407–1440. Retrieved from http://doi.wiley.com/10.1111/jofi.12042 doi: 10.1111/jofi.12042

Dhingra, S., Ottaviaono, G., Sampson, T., & van Reenen, J. (2016). *The Consequences of Brexit for UK Trade and Living Standards*.

Di Giuli, A., & Kostovetsky, L. (2014). Are red or blue companies more likely to go green? Politics and corporate social responsibility. *Journal of Financial Economics*, *111*, 158–180.

Ding, D. K., Ferreira, C., & Wongchoti, U. (2016, oct). Does it pay to be different? Relative CSR and its impact on firm value. *International Review of Financial Analysis*, *47*, 86–98. Retrieved from https://www.sciencedirect.com/science/article/pii/S1057521916301089 doi: 10.1016/J.IRFA.2016.06.013

Duchin, R., Ozbas, O., & Sensoy, B. A. (2010, sep). Costly external finance, corporate investment, and the subprime mortgage credit crisis. *Journal of Financial Economics*, *97*(3), 418–435. Retrieved from https://www.sciencedirect.com/science/article/pii/S0304405X09002566 doi: 10.1016/J.JFINECO.2009.12.008

Dutta, S., & Jog, V. (2009, aug). The long-term performance of acquiring firms: A re-examination of an anomaly. *Journal of Banking & Finance*, *33*(8), 1400–1412. Retrieved from https://www.sciencedirect.com/science/article/pii/S0378426609000302 doi: 10.1016/J.JBANKFIN.2009.02.004

Eccles, R. G., Ioannou, I., & Serafeim, G. (2014). The impact of corporate sustainability on organizational processes and performance. *Management Science*(60), 2835–2857.

Edgington, T. (2019, aug). *Brexit: What trade deals has the UK done so far?* Retrieved from https://www.bbc.co.uk/news/uk-47213842

El Ghoul, S., Guedhami, O., Kwok, C. C. Y., & Mishra, D. R. (2011). Does corporate social responsibility affect the cost of capital? *Journal of Banking & Finance*, *35*(9), 2388–2406. doi: 10.1016/j.jbankfin.2011.02.007M4-Citavi

Erb, C. B., Harvey, C. R., & Viskanta, T. E. (1996, nov). Political Risk, Economic Risk, and Financial Risk. *Financial Analysts Journal*, *52*(6), 29–46. Retrieved from https://www.tandfonline.com/doi/full/10.2469/faj.v52.n6.2038 doi: 10.2469/faj.v52.n6.2038

Fama, E. F., & French, K. R. (1993, feb). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, *33*(1), 3–56. doi: 10.1016/0304-405X(93)90023-5

Fernández-Villaverde, J., Guerrón-Quintana, P., Kuester, K., & Rubio-Ramírez, J. (2015, nov). Fiscal Volatility Shocks and Economic Activity. *American Economic Review*, *105*(11), 3352–3384. Retrieved from http://pubs.aeaweb.org/doi/10.1257/aer.20121236 doi: 10.1257/aer.20121236

Financial Times. (2016). *Brexit poll tracker*. Retrieved 2019-07-12, from https://ig.ft.com/sites/brexit-polling/

Finkelstein, S., & D'Aveni, R. A. (1994, oct). CEO duality as a double-edged sword: How boards of directors balance entrenchment avoidance and unity of command. *Academy of Management Journal*, *37*(5), 1079–1108. Retrieved from http://amj.aom.org/cgi/doi/10.2307/256667 doi: 10.2307/256667

Forum, US Social Investment. (2018). *Report on US Sustainable, Responsible and Impact Investing Trends 2018* (Tech. Rep.). Retrieved from https://www.ussif.org/trends

Francis, B. B., Hasan, I., & Zhu, Y. (2014). Political uncertainty and bank loan contracting. *Journal of Empirical Finance*, *29*, 281–286. doi: 10.1016/j.jempfin.2014.08.004M4-Citavi

Fraser of Allander Institute. (2016). *Long-term Economic Implications of Brexit*. Edinburgh.

Freeman, R. E. (1984). *Strategic management* ([Nachdr.] ed.). Boston, Mass.: Pitman.

Friedman, M. (1970, sep). *The Social Responsibility of Business is to Increase its Profits*.

Gamerschlag, R., Möller, K., & Verbeeten, F. (2011, jul). Determinants of voluntary CSR disclosure: empirical evidence from Germany. *Review of Managerial Science*, *5*(2-3), 233–262. Retrieved from http://link.springer.com/10.1007/s11846-010-0052-3 doi: 10.1007/s11846-010-0052-3

Gao, P., & Qi, Y. (2013). *Political uncertainty and public financing costs: Evidence from U.S. gubernatorial elections and municipal bond markets*.

Gemmill, G. (1992). Political risk and market efficiency: Tests based in British stock and options markets in the 1987 election. *Journal of Banking & Finance*, *16*(1), 211–231. doi: 10.1016/0378-4266(92)90086-FM4-Citavi

Ghoul, S. E., Guedhami, O., & Kim, Y. (2017). Country-level institutions, firm value, and the role of corporate social responsibility initiatives. *Journal of International Business Studies*, *48*(3), 360–385. doi: 10.1057/jibs.2016.4M4-Citavi

Giavazzi, F., & McMahon, M. (2012). Policy uncertainty and household savings. *Review of Economics and Statistics*(94), 517–531.

Godfrey, P. C. (2005). The relationship between corporate philantropy and shareholder wealth: A risk management perspective. *Academy of Management Review*, *30*, 777–798.

Godfrey, P. C., Merrill, C. B., & Hansen, J. M. (2009). The relationship between corporate social responsibility and shareholder value: An empirical test of the risk management hypothesis. *Strategic Managmeent Journal*, *30*, 425–445.

Goodell, J. W., & Vähämaa, S. (2013). US presidential elections and implied volatility: The role of political uncertainty. *Journal of Banking & Finance*, *37*(3), 1108–1117. doi: 10.1016/j.jbankfin.2012.12.001M4-Citavi

Goss, A., & Roberts, G. S. (2011). The impact of corporate social responsibility on the cost of bank loans. *Journal of Banking & Finance*, *35*(7), 1794–1810. doi: 10.1016/j.jbankfin.2010.12.002M4-Citavi

Guenster, N., Bauer, R., Derwall, J., & Koedijk, K. (2011, sep). The Economic Value of Corporate Eco-Efficiency. *European Financial Management*, *17*(4), 679–704. Retrieved from http://doi.wiley.com/10.1111/j.1468-036X.2009.00532.x doi: 10.1111/j.1468-036X.2009.00532.x

Guiso, L., Sapienza, P., & Zingales, L. (2004). The role of social capital in financial development. *American Economic Review*(94), 526–556.

Guiso, L., Sapienza, P., & Zingales, L. (2008, dec). Trusting the Stock Market. *The Journal of Finance*, *63*(6), 2557–2600. Retrieved from http://doi.wiley.com/10.1111/j.1540-6261.2008.01408.x doi: 10.1111/j.1540-6261.2008.01408.x

Guiso, L., Sapienza, P., & Zingales, L. (2015). The value of corporate culture. *Journal of Financial Economics*, *117*(1), 60–76.

Gulen, H., & Ion, M. (2016). Policy Uncertainty and Corporate Investment. *Review of Financial Studies*, *29*(3), 523–564.

Harjoto, M. A., & Jo, H. (2015, apr). Legal vs. Normative CSR: Differential Impact on Analyst Dispersion, Stock Return Volatility, Cost of Capital, and Firm Value. *Journal of Business Ethics*, *128*(1), 1–20. Retrieved from http://link.springer.com/10.1007/s10551-014-2082-2 doi: 10.1007/s10551-014-2082-2

Harty, D. (2019, jul). *House Democrats weigh introducing legislation to mandate ESG disclosure in US*.

Hill, P., Korczak, A., & Korczak, P. (2019). Political uncertainty exposure of individual companies: The case of the Brexit referendum. *Journal of Banking & Finance*, *100*, 58–76. doi: 10.1016/j.jbankfin.2018.12.012M4-Citavi

Hirschey, M. (1985). Market Structure and Market Value. *The Journal of Business*, *58*(1), 89–98.

Hong, H., Kubik, J., & Scheinkman, J. (2012, oct). *Financial Constraints on Corporate Goodness* (Tech. Rep.). Cambridge, MA: National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w18476.pdf doi: 10.3386/w18476

Jens, C. E. (2017). Political uncertainty and investment: Causal evidence from U.S. gubernatorial elections. *Journal of Financial Economics*, *124*(3), 563–579. doi: 10.1016/j.jfineco.2016.01.034M4-Citavi

Jiao, Y. (2010, oct). Stakeholder welfare and firm value. *Journal of Banking & Finance*, *34*(10), 2549–2561. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S0378426610001743 doi: 10.1016/j.jbankfin.2010.04.013

Jo, H., & Harjoto, M. A. (2011). Corporate Governance and Firm Value: The Impact of Corporate Social Responsibility. *Journal of business ethics : JBE*, *103*(3), 351–383. doi: 10.1007/s10551-011-0869-yM4-Citavi

Jo, H., & Harjoto, M. A. (2012). The causal effect of corporate governance on corporate social responsibility. *Journal of business ethics: JBE*(106), 53–72.

Jo, H., & Na, H. (2012, nov). Does CSR Reduce Firm Risk? Evidence from Controversial Industry Sectors. *Journal of Business Ethics*, *110*(4), 441–456. doi: 10.1007/s10551-012-1492-2

Julio, B., & Yook, Y. (2012). Political Uncertainty and Corporate Investment Cycles. *The Journal of Finance*, *67*(1), 45–83.

Kelly, B., Pastor, L., & Veronesi, P. (2016, oct). The Price of Political Uncertainty: Theory and Evidence from the Option Market. *The Journal of Finance*, *71*(5), 2417–2480. Retrieved from http://doi.wiley.com/10.1111/jofi.12406 doi: 10.1111/jofi.12406

Kim, Y., Li, H., & Li, S. (2014). Corporate social responsibility and stock price crash risk. *Journal of Banking & Finance*, *43*, 1–13. doi: 10.1016/j.jbankfin.2014.02.013M4-Citavi

Knack, S., & Keefer, P. (1997). Does social capital have an economic payoff? A cross-country investigation. *Quarterly Journal of Economics*(112), 1251–1288.

Kong, A., Rapach, D. E., Strauss, J. K., & Zhou, G. (2011, jul). Predicting Market Components Out of Sample: Asset Allocation Implications. *The Journal of Portfolio Management*, *37*(4), 29–41. Retrieved from http://jpm.pm-research.com/lookup/doi/10.3905/jpm.2011.37.4.029 doi: 10.3905/jpm.2011.37.4.029

KPMG. (2017). *The KPMG Survey of Corporate Responsibility Reporting 2017*. Retrieved from https://assets.kpmg/content/dam/kpmg/be/pdf/2017/kpmg-survey-of-corporate-responsibility-reporting-2017.pdfY3-28.06.2019M4-Citavi

La Porta, R., Lopez-de Silanes, F., Shleifer, A., & Vishny, R. W. (1997). Trust in Large Organizations. *The American Economic Review*, *87*(2), 333–338.

Lambert, C., Noth, F., & Schüwer, U. (2017, jan). How do insured deposits

affect bank risk? Evidence from the 2008 Emergency Economic Stabilization Act. *Journal of Financial Intermediation*, *29*, 81–102. doi: 10.1016/J.JFI.2015.07.001

Lee, D. D., & Faff, R. W. (2009, may). Corporate Sustainability Performance and Idiosyncratic Risk: A Global Perspective. *Financial Review*, *44*(2), 213–237. Retrieved from http://doi.wiley.com/10.1111/j.1540-6288.2009.00216.x doi: 10.1111/j.1540-6288.2009.00216.x

Li, J., & Born, J. A. (2006). Presidential Election Uncertainty and Common Stock Returns in the United States. *Journal of Financial Research*, *29*(4), 609–622.

Lins, K. V., Servaes, H., & Tamayo, A. (2017, aug). Social Capital, Trust, and Firm Performance: The Value of Corporate Social Responsibility during the Financial Crisis. *The Journal of Finance*, *72*(4), 1785–1824. Retrieved from http://doi.wiley.com/10.1111/jofi.12505 doi: 10.1111/jofi.12505

Lins, K. V., Volpin, P., & Wagner, H. F. (2013, oct). Does Family Control Matter? International Evidence from the 2008–2009 Financial Crisis. *Review of Financial Studies*, *26*(10), 2583–2619. Retrieved from https://academic.oup.com/rfs/article-lookup/doi/10.1093/rfs/hht044 doi: 10.1093/rfs/hht044

Luo, X., & Bhattacharya, C. B. (2009). The debate over doing good: Corporate social performance, strategic marketing levers, and firm-idiosyncratik risk. *Journal of Marketing*, *73*, 198–213.

Makni, R., Francoeur, C., & Bellavance, F. (2009). Causality Between Corporate Social Performance and Financial Performance: Evidence from Canadian Firms. *Journal of Business Ethics : JBE*, *89*(3), 409–422. doi: 10.1007/s10551-008-0007-7M4-Citavi

Minford, P., Gupta, S., Le, V., Mahambare, V., & Xu, Y. (2016). *Should Britain Leave the EU? An Economic Analysis of a Troubled Relationship* (2M4 - Ci ed.). Cheltenham: Edward Elgar.

Minor, D. (2015). *The Value of Corporate Citizenship: Protection.*

Nelling, E., & Webb, E. (2009, feb). Corporate social responsibility and financial performance: the "virtuous circle" revisited. *Review of Quantitative Finance and Accounting*, *32*(2), 197–209. Retrieved from http://link.springer.com/10.1007/s11156-008-0090-y doi: 10.1007/s11156-008-0090-y

Nguyen, L., & Wilson, J. O. S. (2018, dec). How does credit supply react to a natural disaster? Evidence from the Indian Ocean Tsunami. *The European Journal of Finance*, 1–18. Retrieved from https://www.tandfonline.com/doi/full/10.1080/1351847X.2018.1562952 doi: 10.1080/1351847X.2018.1562952

Nguyen, N. H., & Phan, H. V. (2017). Policy Uncertainty and Mergers and Acquisitions. *Journal of Financial and Quantitative Analysis*, *52*(2), 613–644. doi: 10.1017/S0022109017000175M4-Citavi

Orlitzky, M. (2013). Corporate Social Responsibility, Noise, and Stock Market Volatility. *Academy of Management Perspectives*, *27*(3), 238–254. doi: 10.5465/amp.2012.0097

Pantzalis, C., Stangeland, D. A., & Turtle, H. J. (2000). Political elections and the resolution of uncertainty: The international evidence. *Journal of Banking and Finance*, *24*(24 // 10), 1575–1604. doi: 10.1016/S0378-4266(99)00093-XM4-Citavi

Pastor, L., & Veronesi, P. (2012). Uncertainty about government policy and stock prices. *The Journal of Finance*(64), 1219–1264.

Pastor, L., & Veronesi, P. (2013). Political uncertainty and risk premia. *Journal of Financial Economics*(110), 520–545.

Petersen, M. A. (2009, jan). Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches. *Review of Financial Studies*, *22*(1), 435–480. Retrieved from https://academic.oup.com/rfs/article-lookup/doi/10.1093/rfs/hhn053 doi: 10.1093/rfs/hhn053

Phan, H. V., Nguyen, N. H., Nguyen, H. T., & Hegde, S. (2019). Policy uncertainty and firm cash holdings. *Journal of Business Research*(95), 71–82.

Putnam, R. D. (1993). What makes democracy work? *National Civic Review*, *82*(2), 101–107. doi: 10.1002/ncr.4100820204M4-Citavi

Ramiah, V., Pham, H. N. A., & Moosa, I. (2017). The sectoral effects of Brexit on the British economy: early evidence from the reaction of the stock market. *Applied Economics*, *49*(26), 2508–2514.

Rapach, D., & Zhou, G. (2013). Forecasting Stock Returns. In E. Graham & A. Timmermann (Eds.), *Handbook of economic forecasting* (2A ed., pp. 328–383). Elsevier B.V. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/B9780444536839000062 doi: 10.1016/B978-0-444-53683-9.00006-2

Renneboog, L., Ter Horst, J., & Zhang, C. (2008). Socially responsible investments: Institutional aspects, performance, and investor behavior. *Journal of Banking & Finance*, *32*(9), 1723–1742. doi: 10.1016/j.jbankfin.2007.12.039M4-Citavi

Reuters, T. (2019). *Thomson Reuters ESG Scores.* Retrieved from https://www.refinitiv.com/content/dam/marketing/en{_}us/documents/methodology/esg-scores-methodology.pdfY3-29.07.2019M4-Citavi

Rüegg-Stürm, J. (2014). *The new St. Gallen management model*. Houndmills; New York: Palgrave Macmillan.

Sacconi, L., & Degli Antoni, G. (2011). *Social Capital, Corporate Responsibility, Economic Behaviour and Performance*. New York: Palgrave Macmillan M4 - Citavi.

Sampson, T., Dhingra, S., Ottaviaono, G., & van Reenen, J. (2016). *Economists for Brexit: A Critique.*

Sapienza, P., & Zingales, L. (2012). A trust crisis. *International Review of Finance*, *12*, 123–131.

Sapienza, P., Zingales, L., & Jones, S. (2019). *Financial Trust Index.* Retrieved 2019-08-15, from http://www.financialtrustindex.org/contact.htm

Scrivens, K., & Smith, C. (2013). *Four interpretations of social capital: An agenda for measurement.*

Servaes, H., & Tamayo, A. (2013, may). The Impact of Corporate Social Responsibility on Firm Value: The Role of Customer Awareness. *Management Science*, *59*(5), 1045–1061. Retrieved from http://pubsonline.informs.org/doi/abs/10.1287/mnsc.1120.1630 doi: 10.1287/mnsc.1120.1630

Sharfman, M. P., & Fernando, C. S. (2008). Environmental risk management and the cost of capital. *Strategic Management Journal*, *29*(6), 569–592. doi: 10.1002/smj.678M4-Citavi

Shleifer, A., & Vishny, R. W. (1997, jun). A Survey of Corporate Governance. *The Journal of Finance*, *52*(2), 737–783. Retrieved from http://doi.wiley.com/10.1111/j.1540-6261.1997.tb04820.x doi: 10.1111/j.1540-6261.1997.tb04820.x

Sommerlad, J., & Chapman, B. (2019, feb). *Which companies are leaving UK, downsizing or cutting jobs ahead of Brexit?* Retrieved from https://www.independent.co.uk/news/business/news/brexit-companies-leaving-uk-list-job-cuts-eu-no-deal-customs-union-a8792296.html

Stearns, J. (2019, jul). *Von der Leyen Vows $1.1 Trillion Green Deal in Pitch to EU.* Retrieved from https://www.bloomberg.com/news/articles/2019-07-16/von-der-leyen-proposes-green-deal-in-bid-to-clinch-eu-s-top-job

Stock, J. H., & Watson, M. W. (2012). Disentangling the Channels of the 2007–09 Recession. *Brookings Papers on Economic Activity*, 81–135.

Surroca, J., & Tribó, J. A. (2008). Managerial Entrenchment and Corporate Social Performance. *Journal of Business Finance & Accounting*, *35*(5-6), 748–789. doi: 10.1111/j.1468-5957.2008.02090.xM4-Citavi

Thompson, S. B. (2011, jan). Simple formulas for standard errors that cluster by both firm and time. *Journal of Financial Economics*, *99*(1), 1–10. Retrieved from https://www.sciencedirect.com/science/article/pii/S0304405X10001923 doi: 10.1016/J.JFINECO.2010.08.016

Tielmann, A., & Schiereck, D. (2017). Arising borders and the value of logistic companies: Evidence from the Brexit referendum in Great Britain. *Finance Research Letters*, *20*, 22–28.

Ullman, A. E. (1985). Data in Search of a Theory: A Critical Examination of the Relationship's among Social Performance, Social Disclosure and Economic Performance of US Firms. *Academy of Management Review*, *10*, 540–557.

van Essen, M., Engelen, P.-J., & Carney, M. (2013, may). Does "Good" Corporate Governance Help in a Crisis? The Impact of Country- and Firm-Level Governance Mechanisms in the European Financial Crisis. *Corporate Governance: An International Review*, *21*(3), 201–224. Retrieved from http://doi.wiley.com/10.1111/corg.12010 doi: 10.1111/corg.12010

Voth, H.-J. (2002). *Stock price volatility and political uncertainty: Evidence*

from the interwar period.

Waddock, S. A., & Graves, S. B. (1997). The Corporate Social Performance-Financial Performance Link. *Strategic Management Journal*, *18*(4), 303–319.

Waisman, M., Ye, P., & Zhu, Y. (2015). The effect of political uncertainty on the cost of corporate debt. *Journal of Financial Stability*, *16*, 106–117. doi: 10.1016/j.jfs.2015.01.002M4-Citavi

Walker, N. (2019). *Brexit timeline: events leading to the UK's exit from the European Union.* London. Retrieved from https://researchbriefings.parliament.uk/ResearchBriefing/Summary/CBP-7960Y3-14.07.2019M4-Citavi

Wheeler, B. (2017). *Why wasn't 2015 the 'Brexit election'?* Retrieved from https://www.bbc.co.uk/news/uk-politics-39717750

Wilson, J. (2019, aug). *No-deal Brexit odds: Latest predictions on leaving the EU without a deal.*

Xu, N., Chen, Q., Xu, Y., & Chan, K. C. (2016). Political uncertainty and cash holdings: Evidence from China. *Journal of Corporate Finance*, *40*, 276–295. doi: 10.1016/j.jcorpfin.2016.08.007M4-Citavi

Junior Management Science 6(1) (2021) 25-38

Junior Management Science

journal homepage: www.jums.academy

Are Firms Paying for the Minimum Wage? Evidence from Germany

Maximilian Kühn

University of Mannheim

Abstract

Following an intensive discussion, Germany introduced a nationwide minimum wage on 1 January 2015, which was set at an hourly wage of € 8.50. Undeniably, there is a significant amount of international and national literature that discusses the minimum wage, both generally and specifically, in Germany. Counterintuitively, only small, negative employment effects are identified for the German minimum wage, whereas some international studies (e.g. for the US) even found positive effects. However, only a few papers focus on different sorts of adjustment channels that firms are applying as a response to the higher labor costs. This thesis focuses on the analysis of firms' profitabilities and asks if profit margins significantly declined in highly affected industries, as a response to the nationwide minimum wage. Therefore, this paper uses a Difference-in-Difference approach that compares the ratio of pre-tax profits to revenue in industries which are more exposed than their less affected counterparts. To define the exposure of an industry to the minimum wage, the Structure of Earnings Survey of 2014 is used to calculate industry specific bite measures. Surprisingly, no significant decrease in firm's profitability can be found as a response to the German minimum wage introduction in 2015. This result is consistent over all model specifications that are used in the underlying thesis.

Keywords: Minimum wage; Germany; difference-in-difference; adjustment channels; profitability; firm-level performance.

1. Introduction

Since Card (1992) and Card and Krueger (1994) analyses of county minimum wages in New Jersey's fast-food restaurants, which reveal positive employment effects, the classical thinking about minimum wages changed to include a broader range of opinions. Before their study, there was a consensus among economists that a binding minimum wage leads to negative employment effects. However, further research focused on the market specifications and concluded that minimum wages can have positive, negative, or no effect on employment depending on the labor market structure. Many economists, especially after Card and Krueger (1994) study, tried to explain their empirical results with theoretic deviations of the standard model. Inevitably, there is no clear evidence on the different effects of minimum wages, neither on the empirical side, nor on the theoretical side.

Minimum wage research also expanded to employment, its related effects, and possible adjustment channels for firms facing higher labor costs due to the wag floor (e.g. Hirsch, Kaufman, & Zelenska, 2015). The relevance of these studies lies in the ambivalence of employment effects. If firms do not cut jobs but are still confronted with higher labor costs, they might adjust their behavior in terms of their pricing, which would mean that consumers pay for the higher wages, or through lower profits, which would be equivalent to firm owners paying for an introduction of a minimum wage. Thus, only considering employment effects when evaluating minimum wage could result in an incomplete analysis of the situation (Bruttel, Baumann, & Dütsch, 2018).

Prior to 1 January 2015, Germany was one of the few countries in the European Union without a nationwide minimum wage. This changed after two German political parties, the Christian Democratic Party (CDU) and the Social Democratic Party (SPD), agreed on a national wage floor of € 8.50 per hour. This decision came after an intensive debate, where many economists feared massive job loss, especially in Eastern regions of Germany. However, supporters saw a national minimum wage as a way to promote fair labor practices, create a positive distributional effect, and implement an instrument to reduce poverty. In the end, the two primary arguments for wage reform were the increasing gross wage inequality in Germany and the decreasing share of the working population using collective bargaining (Bossler & Gerner, 2016, Bosch, 2015).

In the four years since the German labor market reform,

DOI: https://doi.org/10.5282/jums/v6i1pp25-38

many studies focused on the employment effects. Indeed, these studies only identified small negative employment effects. Such effects included the significant decrease in the prevalence in mini jobs, and firms appear to have postponed their hiring rather than cut full time jobs (Caliendo, Schröder, & Wittbrodt, 2019). Unfortunately, other adjustment channels such as a profit decline have not received much attention in the context of the German minimum wage. Thus, this thesis aims to focus on the German minimum wage's effect on firms' profitabilities, which is defined as the ratio of pre-tax accounting profit to total firm revenue.

The key consideration leading to this analysis is, therefore, that higher labor costs are associated with lower profits if employment does not change. Therefore, it is likely that firms who do not cut jobs, face a lower profitability because of the minimum wage if they are unable to shift increased costs to their consumers through price increases. To identify a possible effect on firms' profitability ratios, this paper applies a Difference-in-Difference (DiD) approach and follows previous studies in the field of minimum wage research (e.g. Mayneris, Poncet, & Zhang, 2018, Harasztosi & Lindner, 2019, Draca, Machin, & Van Reenen, 2011). Two different datasets will be used: the Special Statistical Publication of the German Central Bank (2018), which includes the profit margins on the industry level, and the Structure of Earning Survey (2014) to compute the affectedness for each industry.

This thesis begins with an introduction into the policy context in Germany in section 2, and it will focus on key theoretical considerations on minimum wages and theoretical outcomes on firms' profits in section 3. Part 4 summarizes the empirical literature on international and national evidence, whereas section 5 gives an introduction to the data used. This is followed by a short graphical description of the data in section 6. A detailed description of the empirical strategy is given in part 7 and followed by the empirical results and interpretations in section 8. Finally, conclusions are drawn in section 9.

2. Minimum Wage in Germany: Policy context

After the agreement of the Grand Coalition, consisting of the SPD and the CDU, a statutory national minimum wage of € 8.50 per hour was introduced on January 1, 2015. Prior to 2015, only sectoral minimum wages existed, such as for hairdressers and security services (Caliendo, Fedorets, Preuß, Schröder, & Wittbrodt, 2017). Surprisingly, not all unions were in favor of the implementation of a national minimum wage and not all firms were against it. However, employer associations commonly argued against a nationwide minimum wage, as they said it would weaken the German system of collective bargaining. Nevertheless, these cases show the diversity of opinions and the divided interests among all actors in the debate (Bosch, 2015).

Normally, companies in Germany set wages through collective bargaining, where unions negotiate wages with employer associations freely and without any intervention from the government. However, the fraction of people covered by collective bargaining has been decreasing for more than 20 years. One potential reason is the voluntary participation in the collective bargaining process. Additionally, as collective bargaining coverage fell, real wages stagnated, and the wage inequality rose. Combined with the good form of the German labor market, this potentially influenced the decision to introduce the national statutory minimum wage. Furthermore, it was the German chancellor, Angela Merkel, herself who liked the idea of a minimum wage. This was first mentioned by Merkel in 2013 and led to a conflict with the coalition partner—the German liberal party (FDP)—at that time (Bossler & Gerner, 2016).

The Minimum Wage Commission (Mindestlohnkommission, 2016b) proposed further adjustments; on January 1, 2017, the minimum wage increased to € 8.84 per hour, with a future increase to € 9.19 in 2019 and € 9.34 by 2020. In 2014, around 5.5 million workers earned less than € 8.50 per hour. However, only 4 million of these workers were eligible for the nationwide minimum wage according to the Federal Statistics Agency. Most of those who did qualify were workers in the eastern regions of Germany, women, and marginal employed workers ("minijobbers") (Destatis, 2016). However, the introduction of a minimum wage came with many exceptions, and therefore, not all workers who earn less than € 8.50 were eligible. To make finding employment easier, long-term unemployed persons are allowed to work for the first six months for a wage that is below € 8.50. Another notable exception includes interns and workers under the age of 18. This exception does not extend to apprenticeship though, as the Grand Coalition introduced a separate general minimum wage for apprentices of € 515 per month in May 2019. This will be effective for new contracts as of January 1, 2020, even though this only applies to branches where no sectoral agreements exist (Bundesregierung, 2019).

As of 2019, 22 of the 28 countries in the European Union have a nationwide minimum wage. The range is quite high, as can be seen by the stark difference between the Bulgarian minimum wage of € 1.72 per hour compared to Luxemburg's € 11.97. While the German minimum wage of € 8.50 falls between these extremes, a further increase of Germany's minimum wage is highly discussed (Handelsblatt.com, 2019) as the current level of € 9.19 is considered as relatively high (Caliendo et al., 2019).

3. Theoretical Considerations

The following section of this paper will give a short overview of the most relevant labor market theories, with a particular focus on the effects of a minimum wage on market equilibrium.

The oldest and one of the most discussed theories in economics is the case of a perfectly competitive market—the neoclassical theory. In terms of labor markets, this means that firms and consumers do not have any market power; thus, the equilibrium wage w^* and employment level L^* are defined through labor demand and labor supply. Introducing a minimum wage $w_{min} > w^*$ leads to an excess of labor

supply and, therefore, to unemployment. This can be easily seen in the simplest economic model: due to an increase in wages (the price of the product labor), demand for that product falls as labor demand is a downward sloping function. Additionally, total output in the economy falls (Krasniqi, 2007). The full coverage of all workers by the minimum wage is an important underlying assumption in the neoclassical model. Stigler (1946) argues that unemployment might not rise when low-wage workers increase their productivity so that their Marginal Product of Labor (MPL) is at least as high as the introduced minimum wage (Stigler, 1946; Brown, Gilroy, & Kohen, 1982). In addition, the introduction of a binding minimum wage does not affect the level of employment. A profit-maximizing firm will increase their prices in response to the higher labor costs. As a consequence, both the firm's output and its profit decrease (Neumark & Wascher, 2008). Nonetheless, there are some expansions of the standard model, which allow for partial coverage only (e.g. Welch, 1974; Gramlich, 1976) or introduce the case of a heterogeneous worker-force—namely by distinguishing between skilled- and unskilled workers.

However, the result is different if employers have market power and the labor market is characterized by a monopsony. It was Robinson (1933) and Stigler (1946) who first established this idea. A monopsonistic labor market is characterized by only one employer (monopson) whereas there are many possible workers. Therefore, the firm is no longer a price-taker, and thus has market power, at least over the labor market. Such a firm will find a wage w_M optimal, which is smaller than the equilibrium wage w^* in a perfectly competitive labor market. Firms set their wage optimal when their marginal cost of labor equals their marginal revenue product of labor. Similar to the monopoly case, the marginal cost of labor is higher than the labor supply—easier said, the marginal cost curve is steeper than the labor supply curve. Therefore, an introduction of a minimum wage can stimulate employment under certain circumstances, namely when the following equation holds true: $w_M < w_{min} \leq w^*$

This is because marginal costs decrease, even if average costs rise (Robinson, 1933). Nevertheless, a minimum wage that is set too high clearly predicts the same outcome in the monopsonist model as the neoclassical does. Furthermore, Stigler notes that the application of the monopsony model to labor markets is not "relevant" (Stigler, 1946, p.360). One reason is that low-wage industries are competitive (Stigler, 1946). Additionally, a firm's profit stays unchanged in the monopsony model—compared to the standard neoclassical model (Neumark & Wascher, 2008

Bhaskar and To (1999) try to simulate the effects of a minimum wage in a monopsonistic competition model. Contrary to the monopsony model, the situation of monopsonistic competition has a large number of employers (firms), and they can freely enter or exit the market. These employers all compete with each other for workers. Bhaskar and To (1999) use horizontal differentiation to allow workers to have different preferences on non-wage characteristics to then apply the spatial model of Salop (1979), which originally stems from industrial economics, and models the location, the number of firms, and the prices in equilibrium. If labor markets can be described as monopsonistic competitive, a minimum wage leads to a rise in employment in individual firms. However, some firms will exit due to the negative effect on their profits because of the wage-setting externality; thus, no clear effect on aggregate industry employment can be determined in the long-run, as the number of firms that exit stays ambiguous. Bhaskar and To (1999) argue a highly distorted labor market might increase the economy wide employment.

Another well-known theory in labor markets is the so-called efficiency wage theory, first developed by Shapiro and Stiglitz (1984). They argue that it can be optimal for firms to pay a higher wage (efficiency wage) than the competitive market wage to ensure that workers make an effort and are productive. This model is constructed on the basic principal-agent problem, which means that firms cannot control whether workers work efficiently (Shapiro & Stiglitz, 1984). Rebitzer and Taylor (1995) build on that model and showed that, under certain assumptions, an introduction of a minimum wage can increase short-term employment in the context of an efficiency wage model. An increase or an introduction of a minimum wage, which is higher than w^*, makes a job loss harder. Therefore, the productivity of the workers increases, and in the context of an efficiency model, less resources for production oversight are necessary. Consequently, firms can hire additional workers instead of using their resources to monitor their existing employees. In the long-term, they show that employment can increase, fall, or stay unchanged as a consequence of the firm's long-term zero economic profit condition. This is due to the decrease in firms' profits in their theoretical model. In addition, they argue, that due to the envelope theorem, a minimum wage might even be pareto-improving (Rebitzer & Taylor, 1995). In general, firms' profits do not change due to the minimum wage in efficiency wage models (Neumark & Wascher, 2008). However, in models where firms do not maximize their profits, it is even possible to experience an—ceteris paribus—increase in profitability as a consequence of the minimum wage. This can be explained with the cost saving efforts that companies are forced to pursue because of the higher labor costs and the incentive to operate more efficiently; thus, there may be a situation where an enterprise's cost savings exceed its higher labor costs (Neumark & Wascher, 2008).

To sum up the theoretical part of this thesis, the effect of minimum wages highly depends on the underlying market structure and the specific assumptions made. However, this means that there is no consensus about a generic effect of minimum wages among economists, and therefore, this "issue is best approached empirically through the careful study of the specifics of particular minimum wage laws and the operation of particular labor markets" (Rebitzer & Taylor, 1995, p. 254).

4. Empirical findings so far: Literature overview

In this section, a short description about empirical studies of the effects of a minimum wage will be given—with a special focus on, but not limited to, the German minimum wage.

4.1. Minimum Wage studies: International

Surprisingly, not only are the theoretical aspects and the possible impacts of a statutory minimum wage highly ambiguous and not clearly determined, but the empirical research also rarely agrees on a common conclusion. Nonetheless, until the work of Card and Krueger (1994) and the rising New Minimum Wage Research, economists mainly supported the concept of minimum wage causing a disemployment effect. This is especially due to the review of the existing research so far by Brown et al. (1982), where they study the literature concerning a minimum wage in the US. Brown et al. (1982) find disemployment effects, though they emphasized the rather small consequence, mainly for teenagers or the youth. Additionally, they identified an elasticity of teenage employment with respect to the minimum wage of -0.1 to -0.3; thus, it can be concluded that a 10% increase in the minimum wage leads to a reduction in total teenage employment by 1-3% (Brown et al., 1982).

However, since the work of Card and Krueger (1994); Card and Krueger (1995); Card and Krueger (2000), some economists now challenge the traditional neoclassic understanding of minimum wages and labor markets. Card & Krueger, 1994 analyze the effect of the increase of New Jersey's minimum wage in 1992 from $4.25 to $5.05. For that purpose, they survey 410 fast-food restaurants in New Jersey and the neighboring state of Pennsylvania before and after this increase. They were the first to apply the method of a difference-in-difference estimation that uses a regional variation in the affectedness to the study of minimum wage effects (Caliendo et al., 2019). To illustrate, they compare the difference in employment between New Jersey and Pennsylvania before and after the exogeneous rise in New Jerseys binding minimum wage. The result was revolutionary. They not only found no significant disemployment effects for New Jersey's fast-food restaurants relative to their neighboring state, but they also found slightly positive employment effects. As compared to restaurants in Pennsylvania, employment in New Jersey rose by 13% (Card & Krueger, 1994; Card & Krueger, 1995; Schmitt, 2013). Neumark and Wascher (2000) reply to Card and Krueger (1994) evaluation of New Jersey's higher minimum wage by using payroll data. They find significant negative employment effects and conclude by emphasizing the consistence of their results with the general labor market theories that predict a decrease in employment—especially for workers at the end of the wage distribution (Neumark & Wascher, 2000).

In contrast to Card (1992), Card and Krueger (1994), Card and Krueger (1995) and to Neumark and Wascher (2000), Michl (2000) argues that both can be right due to rescheduling. That's because Card and Krueger (1994) define employment as the total number of workers, an increasing value, whereas Neumark and Wascher (2000) analyze payroll data and define employment as the total payroll hours, a decreasing value. According to the rescheduling model of Michl (2000), a decrease in weekly hours is consistent with a small, though not significant, rise in total workers. Summing up, workers are better off after the minimum wage since the increase in wages exceedes the decline in working hours (Michl, 2000). Other recent studies stress the importance of a proper design of the control group, and thus, many analyses lead to different and often contradictorily results due to differences in the applied method. Namely, the disemployment effects might be a consequence of an omitted variable bias, which means that some studies fail to control for effects which influence the underlying dependent variable (Dube, Lester, & Reich, 2010). When controlling for spatial heterogeneity in employment growth, Dube et al. (2010) cannot find any employment effects as a consequence of the minimum wage. Allegretto, Dube, and Reich (2011) support that finding by controlling for long-term growth differences among states and heterogenous economic shades. Their study emphasizes the non-existence of negative employment effects; instead, they explain that earlier studies with negative results use biased and non-robust estimators (Allegretto et al., 2011).

Even though there are numerous studies focusing on the employment effects of minimum wages, there is a lack of research that deals with other possible adjustment channels. Firms' profits are one such channel. Draca et al. (2011) analyze the effect on a firm's profitability after the introduction of a nationwide minimum wage in the United Kingdom (UK) in 1999. To perform these studies, the authors look at two different datasets. On the one hand, they analyze the home care sector, where they expect a rather significant effect. On the other hand, the authors look at accounting data from a subsample of all firms that are registered in the UK. They then compare firms' profit margins—both highly affected as well as non-affected—before and after the introduction of the minimum wage. In the home care sector, firms' profit margins, defined as the ratio of profit to revenue, decreased by 23%. Whereas in the sector wide data, the average profit margin decreased in the range of 7.8% to 10.7%. However, they were not able to find evidence for a higher probability of market exit or a higher productivity due to the minimum wage (Draca et al., 2011).

In addition, study by Harasztosi and Lindner (2019) analyzes the increase in Hungaria's minimum wage in 2001—focusing on several possible adjustment channels such as prices and firms profitabilities. They use a difference-in-difference approach to compare both a firm's behavior towards highly affected establishments as well as the behavior of their less-affected counterparts. In line with several other studies, their analysis could only identify a small negative effect on employment; however, they also saw a large wage increase of approximately 50%[1]. Most interestingly, Harasztosi and

[1] Only workers who did not lose their jobs had on average a wage increase

Lindner (2019) concluded that nearly 80% of employee wage increases is paid by the consumers through higher prices. The remaining 20% pay the firm's owners through lower profits[2] due to higher labor costs. Consequently, the increase in the Hungarian minimum wage has a stronger effect on industries that cannot easily adjust their prices as a response to higher costs. To be more precise, the profit decline of 0.7% in a firm's sales in 2000, in the short-term, is statistically significant at the 1% level. Regardless, the profit decline of 0.7% only relates to firms highly affected by the minimum wage increase. The long-term effect was actually much weaker. Harasztosi and Linder conclude their research by noting that the minimum wage increase in Hungary is paid by consumers, and thus their study mainly excludes the adjustment channel through lower firm profits. By combining their key results, such as no strong disemployment effects, higher prices, and no stark decline in firm's profitability, they raise the question of whether a minimum wage can therefore redistribute income from consumers (who now pay higher prices) to low-wage employees.

Mayneris et al. (2018) studied the large increase to minimum wages in China in 2004—especially with regards to firms' adjustments behavior. For this purpose, Mayneris et al. (2018) use balance sheet data from the National Bureau of Statistics as well as the China Labor Net dataset for information dealing with the level of various minimum wages. Mayneris et al. (2018) then compare the results of both exposed and non-exposed firms before and after the minimum wage reform. A firm is considered exposed if the average wage is below the local minimum wage. Their results are in line with former studies in many regards, but their findings also present key contradictions. At first, they identify a significant wage increase in exposed firms, but there is no effect on employment. These results are in line with several other papers (e.g. Harasztosi & Lindner, 2019, Draca et al., 2011). In contrast to that, Mayneris et al. (2018) do not only identify no change in firms' profits, but they also detect a significant increase in firms' productivity. Unlike Draca et al. (2011), Mayneris et al. (2018) research finds a greater probability for market exit in highly affected firms. Thus, Mayneris et al. (2018) argue that firms adjust for the higher labor costs with higher productivity, therefore, the level of employment and the firm's profitability does not change. This higher productivity is mostly explained by better inventory management and an increasing capital-labor ratio (Mayneris et al., 2018). The results might be driven by the specific characteristics of China, given that it is not a fully developed country and the large improvement in productivity might be especially relevant for similar-developing economies (Mayneris et al., 2018). Consequently, these findings need to be interpreted very carefully, as their argumentation channel will probably not hold for every industrialized country.

of 50%

[2]Here profits are defined as the following accounting number: Earnings before Iterests and Taxes (EBIT)

4.2. Minimum Wage studies: Germany

Bossler and Gerner (2016) were the first to analyze the effect of the German nationwide minimum wage. They apply a Difference-in-Difference approach on the firm level. Namely, they look at highly affected firms (the treatment group) and compare the results with less-affected establishments (the control group). Their paper finds a decrease in employment of 1.9% in affected firms and implies a small negative elasticity with respect to wages.

Garloff (2016) uses regional differences of the effect (the bite) of a minimum wage to analyze employment. In his study, he detects a significant negative effect on marginal employment ("mini jobs"). Nonetheless, Garloff (2016) identifies a significant positive effect on employment for socially insured jobs. Therefore, employment structure changes, which is in line with Vom Berge and Weber (2017), who also find a decrease in the number of mini jobs. To illustrate, between 150,000 and 200,000 mini jobs were lost due to the minimum wage (Schmitz, 2017). Therefore, claim to identify a restructuring process on the labor market from marginal employment to jobs with social insurance. Though, many studies find evidence for a decline of mini jobs or a rise in employment (e.g. Garloff, 2016, Schmitz, 2017, Bruckmeier & Becker, 2018, Vom Berge & Weber, 2017, Holtemöller & Pohle, 2017), not all academics find causal evidence that this change was induced by the minimum wage. Holtemöller and Pohle (2017) do not support—according to the results of their analysis—the argumentation of a minimum wage induced changed in the labor market.

Some analyses find that the German minimum wage led to a decrease in (contractual) working hours and an increase in low-wage workers' hourly wages (Caliendo et al., 2017, Caliendo, Fedorets, & Schröder, 2018, Bonin et al., 2018, Bossler & Gerner, 2016). Nevertheless, Caliendo et al. (2017), Caliendo et al., 2018) could not identify any significant effect on low-wage workers' monthly salaries. Another recent paper uses differences in the regional bite of the minimum wage and then applies a Difference-in-Difference approach. It finds that an increase in the regional bite by 1 percentage point leads to a 0.5% higher growth in hourly wages at the lower end of the wage distribution (Ahlfeldt, Roth, & Seidel, 2018). However, most studies only identify modest employment effects and find nearly no negative effects. Others emphasize firms' strategy to decrease hiring instead of directly cutting jobs in the short-run. Therefore, the studies mostly associate the negative employment with a lower growth in employment (Bossler & Gerner, 2016, Bossler et al., 2018).

Counter intuitively, Bruckmeier and Becker (2018) argue that there is no minimum wage induced change of the probability to find a job for the unemployed. Before the introduction of the minimum wage, many opponents argued that a wage floor would disadvantage the unemployed, as their qualifications do not match the relatively high minimum wage of € 8.50. According to another paper, there might even be a potential increase in the reservation wages of unemployed persons (Fedorets & Filatov, 2018).

Many studies fail to identify significant negative employment effects. However, as firms' labor costs increase, companies will adjust their behavior. A possible adjustment channel is to increase prices; Link (2019) shows that affected firms in Germany increase their prices more frequently compared to their less-affected counterparts. However, this result is not entirely limited to low-wage industries. Interestingly, firm adjustment through higher prices questions the description of the German labor market as a monopsony (Link, 2019). Bossler et al. (2018) use data from the Institute for Employment research (henceforth: IAB) to analyze the effects of the introduced German minimum wage on the firm level to identify certain adjustment channels of German firms. They use data from 2011 to 2016 to look at possible effects on productivity or profitability. In their study, productivity is measured with revenue and the numbers of employees; easier said revenue per worker is looked at. However, Bossler et al. (2018) cannot find a significant difference in productivity between highly affected and less-affected firms. Nevertheless, they show that more exposed firms to the minimum wage have a ca. 2% higher probability to make accounting losses than their non-affected counterparts. Additionally, they identify a significant, though a rather weak, effect of the minimum wage on a firm's profitability due to the higher labor costs. Nonetheless, Bossler et al. (2018) were not able to find clear evidence for differences in firm closures.

5. Data description

In the underlying thesis, two different datasets will be used. The first dataset is the Special Statistical Publication Number 6 (henceforth: SSP) of the German Central Bank (Statistische Sonderveröffentlichung 6, Deutsche Bundesbank). It contains ratios of financial statements of German enterprises. This study is used to derive the profit margins of its reflected sectors at the aggregated level. The profit margin here is defined as the ratio of Pre-tax profit (Earnings before Taxes; EBT) to total revenue. In 2016, the SSP contained data of 116,000 balance sheets. However, the dataset itself does not contain all 116,000 financial reports to avoid an influence on the data due to a sample effect. Therefore, only 83,000 balance sheets are used to analyze firms' financial statements. The data provided by the German Central Bank clearly faces some pitfalls that should be considered—especially when drawing conclusions from the results. First, it contains financial accounting statements of many large corporations, but Small and Medium Enterprises (SME) are underrepresented. For example, firms with an overall revenue higher than € 50 million make 86% of the dataset compared to 64% of the statistical population according to the German business register. This could present a problem, as small and medium sized firms are typically more affected by a nationwide minimum wage compared to large, multinational corporations. Second, no good comparison of West and East Germany[3] is possible due to a lack of specific results for the East, as the dataset inadequate regional industry records to get representative results. Finally, relying on accounting data is always risky, as some results may possibly reflect a change in a company's accounting policy or might be the result of certain tax improving actions. To account for that, pre-tax profits (EBT) will be used in this thesis (Deutsche Bundesbank, 2018).

Nevertheless, the German Central Bank's SSP reveals some benefits as well. Namely, the SSP follows the WZ industry classification (2008) of the Federal Statistics Agency (FSA) (Destatis, 2008). Additionally, it provides relative ratios with a high degree of details regarding sub-sector classifications. The use of ratios instead of absolute values will give more representative results, as ratios are not affected by macroeconomic factors such as inflation. Furthermore, considering only an absolute increase in profits reveals no information concerning the company's profitability. Therefore, the use of a profitability ratio may probably show more realistic results. The second dataset is the Structure of Earnings Survey SES (Verdienststrukturerhebung VSE) of Germany's Federal Statistics Agency from 2014. This statistic contains survey data about individual earnings. It takes place every 4 years, is compulsory, and contains data from 60,000 firms and 1,000,000 employees in 2014. In addition, the SES follows the WZ industry classification of 2008 as well (Destatis, 2019). This thesis uses calculated bites of the German minimum wage in the different industries. These bites were calculated using the Structure of Earnings Survey dataset.

6. Summary statistics

This section gives a short overview over the empirical statistics of the SSP dataset which was introduced in section 5. Figure 1 shows the development of firms' profit margins[4] over four years prior to and two years following the introduction of Germany's minimum wage. The graph illustrates the average profit margins, aggregated on the industry level. By looking at 2015, the main year of interest, a clear decline in profit margins compared to the year 2014 can be seen. This is the strongest decline for the time period of 2010 to 2017. Nevertheless, there has been a decline from 2012 to 2013, where no minimum wage was introduced. Additionally, the level of profitability in 2015 is still at a higher level, approximately 5.0%, than in the years 2010 to 2013. Furthermore, profit margins increase for the post-reform years 2016 and 2017. Figure 2 gives a deeper insight into the differences of exposed and non-exposed industries' profit margins. This graph uses the binary definition of the bite, which means that an industry is considered exposed if its bite exceeds the average share of eligible workers (approximately 10.9%). The

[3] Note that the terms West and East Germany refer to the former national territory of the Federal Republic of Germany (West) and the German Democratic Republic (East) until 1990.

[4] Profit margin as Earnings before Tax to total Revenue ratio (in percentages)

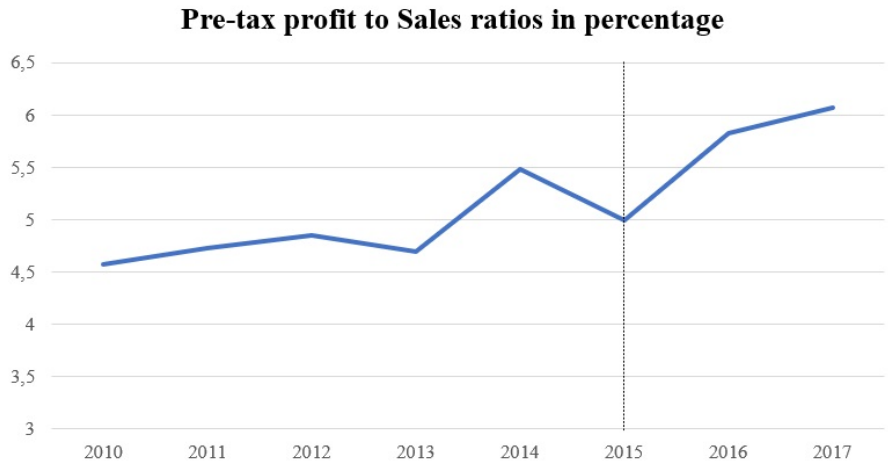


Figure 1: Industry's profit margin development

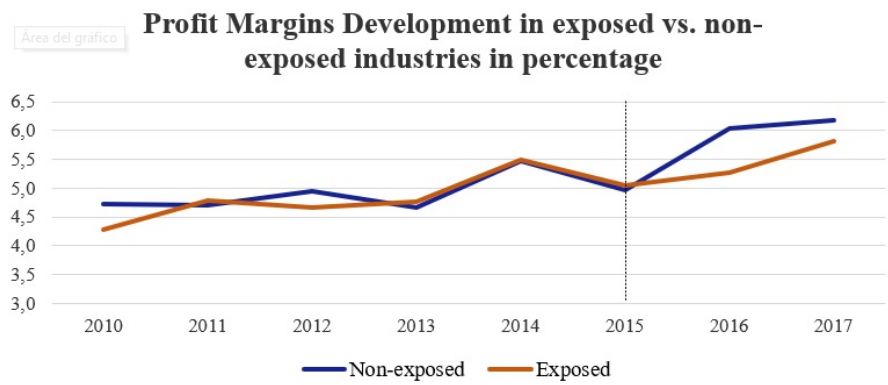


Figure 2: Exposed vs. non exposed industries profit margin development

orange line represents the profit margins of firms in exposed industries, and the blue line provides information regarding the non-exposed industries. Similar to Figure 1, there is a clear decline in profitability for both groups in 2015. However, the expectation that firms in exposed industries face a stronger decline is not met. Indeed, the level of exposed and non-exposed profitability is nearly the same for 2015. Only the further development for the years 2016 and 2017 shows a smaller increase in profitability in exposed firms compared to those which are not affected by the labor market reform of 2015. For the years 2013 and 2014, both groups were on a common trend as their profit margins nearly equal each other, though there are small differences from 2010 to 2012. Furthermore, Figure 2 illustrates that since the minimum wage reform, both groups have developed with a greater dispersion than before.

7. Empirical Strategy

7.1. General Strategy: The Difference-in-Difference approach

Due to the difficulty in exploring the effect on firms' pre-tax profit to revenue margins that can be causally interpreted as a result of the minimum wage, this analysis applies a quasi-experiment approach. The Difference-in-Difference (DiD) approach compares firms profits in highly affected industries with their less-affected counterparts before and after the introduction of the German nationwide minimum wage. Thus, this analysis relies on defining a treatment group, that includes industries which are highly exposed to a minimum wage, and a control group, that includes industries which are not at all or only slightly affected by this specific policy reform (Draca et al., 2011). Following Caliendo et al. (2019), this study relies on the assumption that a higher affectedness to the minimum wage implies a higher effect on the wages and, therefore, the variable of interest in this paper—the profitability.

The SES dataset defines the treatment and control groups, which allows for the calculation of the share of workers who earned less than the introduced wage floor of € 8.50 on an industry level. This share of eligible workers for the minimum wage in a specific industry is used for the "bite measure" and can be seen as the level of affectedness an industry faced.

The analysis uses two different definitions of the bite. The

first definition considers the bite as a continuous variable following recent studies (see e.g. Draca et al., 2011, Harasztosi & Lindner, 2019, Link, 2019, Fedorets & Filatov, 2018). Therefore, the continuous bite measure is very straight forward and only reflected by the share of eligible workers for the minimum wage in industry i. The second definition follows the traditional DiD approach and uses a binary bite measure. To obtain a binary variable of the exposure of an industry, this thesis follows previous studies such as Schmitz (2017) or Caliendo et al. (2017) to calculate the average share of workers eligible for the minimum wage over all industries that are available in the 2014 SES survey. Industries with a treatment intensity that equals or exceeds the average bite are considered as "exposed" to the minimum wage and are part of the treatment group; industries with a bite below the average treatment intensity form the control group and are considered "not exposed". Instead of using regional differences in the bite like Schmitz (2017), this analysis uses industry differences in the treatment intensity as an exogeneous variation. The DiD approach is applied by using specifications of the equation in (1) and (2) which follows the previous studies of Bossler et al. (2018).

$$Y_{it} = \beta_1 \times FA_i + \beta_2 \times Dummy_t^{2015} + \delta(FA_i \times Dummy_t^{2015}) + \theta_i + \gamma_t + \varepsilon_{it} \tag{1}$$

$$Y_{it} = \beta_1 \times Exp_i + \beta_2 \times Dummy_t^{2015} + \delta(Exp_i \times Dummy_t^{2015}) + \theta_i + \gamma_t + \varepsilon_{it} \tag{2}$$

Equation (1) measures the effect on the outcome variable Y_{it} with a continuous bite, the fraction affected variable (FA_i). In equation (1), the calculated share of workers earning less than the minimum wage is used as treatment intensity. Contrarily, the regression in equation (2) is based on the binary definition of the treatment and control group, meaning Exp_i, standing for exposed, is a binary variable that equals 1 if the treatment intensity in industry i exceeds the average treatment intensity and is 0 otherwise. In the analysis of this paper, the outcome variable Y_{it} refers to the EBT to Revenue ratio. However, the interpretations of the coefficients do not change in (1) and (2). $Dummy_t^{2015}$ is a dummy variable which is 1 for every time period after 2015, whereas 2015 is included. The use of FA_i or Exp_i as explanatory variable controls for differences among industries that already existed before the introduction of the nationwide wage floor. Additionally, including the dummy to reflect the fact that if data is pre- or post-reform measures industry-independent time effects. Besides, a significant negative coefficient of the dummy would mean that regardless of the industry, profit margins decreased for years after the policy reform. To estimate the causal effect, the coefficient δ of the interaction term between the treatment intensity and the time dummy is of main interest in the DiD analysis. This coefficient shows how the dependent variable Y_{it}, here the profit margin, developed differently over the time of the underlying panel dataset in highly affected (in terms of the continuous bite definition) or exposed (in terms of the binary bite definition) industries, compared to their less affected or non-exposed counterparts. Thus, this difference is causally interpreted as a response to the introduction of the minimum wage (Bossler et al., 2018, Stock & Watson, 2015).

Unfortunately, due to the aggregation of the data used in this thesis, controlling for individual firm-specific characteristics or other individual effects is hardly possible. To account for this fact, this paper uses the fixed effects model (FE) as a common model in panel data analysis. Furthermore, it uses a random model (RE) as well as a time fixed effects (TFE) model. This is done for two main reasons: first, using a FE and TFE model specification partly accounts for the fact that no other control variables are included. Second, it allows for the identification of differences in the result depending on which model specification is used. Thus, if the results are similar, it is more likely that they do not depend on the specific model used but, instead, are rather robust.

The FE model accounts for a common problem in regression analysis: the omitted variable bias (OVB) problem that is caused when a variable in the error term is correlated with an explanatory variable and explains the dependent variable. To account for this problem, control variables are normally included. The FE model in this underlying panel data regression model, however, allows for correlation between unobserved characteristics about the industries that do not change over time therefore, they are "fixed". To specify, θ_i in the above stated equations measures the industry specific effects which do not vary over time. The TFE model is similar to the FE model, whereas it controls for unobserved heterogeneity that changes over time, but it does not vary across industries. The above stated equations represent a fixed effects model with time fixed effects as well; however, this thesis will present the results for various specifications of these equations, such as only a FE model, a random effects (RE) model that does neither include θ_i nor γ_t, as well as the FE and RE model with time fixed effects (Stock & Watson, 2015).

7.2. Assumptions

As in every economic model, applying the Difference-in-Difference approach underlies some key assumptions. Following Schmitz (2017) and Bossler et al. (2018), there are three main assumptions that must apply to be able to interpret the results as causal effects: Common Trend Assumption, Stable Unit Treatment Value Assumption, and Definition of Treatment and Control Groups.

a) Common Trend Assumption (CTA) The CTA means, that in an absence of a certain policy reform, both groups would have developed similarly. Therefore, if a general nationwide German minimum wage had not been introduced in 2015, the pre-tax profit to revenue ratio would have been on a parallel trend across

highly affected and less-affected industries (Schmitz, 2017). Thus, an identified difference can be causally interpreted as a result of the policy reform—the minimum wage—that was introduced. Due to the fact that the DiD approach is a quasi-experiment, this assumption can unfortunately not be directly tested. As the minimum wage was introduced, it is not possible to determine what would have happened without this specific labor market reform. However, one might look at pre-minimum wage data to determine how profit margins developed in the treatment and the control group (Schmitz, 2017). Additionally, the CTA is often tested by applying placebo regressions. This study follows previous research in the minimum wage field that apply a DiD approach (e.g. Mayneris et al., 2018, Harasztosi & Lindner, 2019, Draca et al., 2011, Garloff, 2016) and conducts placebo regressions in section 9 to reduce the possibility of a violation of the Common Trend Assumption.

b) Stable Unit Treatment Value Assumption (SUTVA) The SUTVA generally means that there are no interdependencies between the treatment and the control group. In other words, the minimum wage only affects the treatment group. Furthermore, the assumption for the use of a continuous bite measure is slightly different as it requires industries with a smaller bite to be proportionally less affected (Schmitz, 2017). This assumption is not as trivial as it seems and can easily be violated in the medium to the long-term, namely through spillover effects that cause interdependencies between the treatment and the control group (Bossler et al., 2018, Bossler & Gerner, 2016). Assuming affected and non-affected industries are not generally related, a policy reform might as well affect the control group. Assuming firms with a very high treatment intensity (bite measure) increase their prices[5] to account for higher labor costs and the end market is competitive, this price increase also affects firms that are part of the control group. Due to the price increase of their competitors, non-affected firms are likely to sell more of their products if consumers are able to substitute the products of an exposed firm with similar goods of a non-exposed firm. Through spillover effects, this specific non-exposed firm now meets a higher demand than in the absence of the minimum wage; thus, there might be a situation where firms in the control group are still affected by the minimum wage. This example might be generally possible, but it is unlikely to happen in the model specification of this paper due to the fact that the analyzed data is highly aggregated and on the industry level. Substituting the same product with that of a firm in another industry is not realistic. Another example for a positive spillover effect that follows Bossler et al. (2018) would be the possibility of an effect on the control group if treated firms react on the minimum wage with a lower output and job cuts.

   Nevertheless, the SUTV assumption is unlikely to be violated in the study of this thesis due to two main reasons. First, these spillover effects need time and do not play a remarkable role in the short-term analysis of this thesis (Schmitz, 2017). Second, previous research about the German minimum wage did not find wage spillovers, which at least excludes the probability of spillovers through wages (Caliendo et al., 2017, Link, 2019).

c) Definition of the treatment and the control group This assumption requires a proper definition of the treatment and the control groups to avoid an estimation bias in the variable of interest. That is specifically relevant as the coefficient of interest in this study—DiD—is the interaction term of the treatment intensity and the time dummy. An incorrect estimation of the treatment intensity would lead to a bias in the coefficient of the interaction term, which is of our main interest in the Difference-in-Difference approach (Bossler et al., 2018).

Such a misleading definition of the two groups can be caused by an anticipating behavior of firms (Bossler et al., 2018). This might be the case if firms increase their wages before the 1 January 2015, as firms know that they have to pay higher wages to their employees who currently earn less than € 8.50. However, this can cause a problem if these firms are treated as control firms only because they are paying wages equivalent or exceeding the introduced wage floor. Treating them as a control group can lead to a wrong definition of the treatment and the control groups as they would not have been part of the control group without their anticipating behavior. This means that firms with the characteristics of low wage firms will be considered as not affected only because they anticipated their high degree of affectedness. To account for that phenomena, Bossler et al. (2018) identify firms with an anticipating behavior and exclude them from their analysis (Bossler et al., 2018). Unfortunately, this is rarely possible in the underlying study. Since data is only available on the industry level and not on the level of individual establishment, this paper does not account for anticipating effects.

Additionally, using a binary bite measure may result in a wrong definition of the control and treatment group. Namely, an exposed industry will include firms that are not affected by the minimum wage as well as some affected firms are part of non-exposed industries and thus are in the control group (Mayneris et al., 2018). To account for that fact, the estimations in this paper present the results for the same regressions as in (1), which only differ in the definition of the bite measure (2).

[5] See for example Harasztosi and Lindner (2019) who find a significant increase in prices for consumers in their analysis of the Hungarian labor market reform. Link (2019) showed that in the German case, more affected firms are more likely to increase their prices. Thus, the above stated assumption is likely to be met.

8. Regression Results and Interpretation

8.1. Results

This section presents the results for the different specifications of equation (1) and (2) of section 7. For both definitions of the treatment intensity, the results of the four different models FE, RE, FE with time fixed effects, and RE with time fixed effects are presented in Tables 1 and 2.

Looking at the results in the continuous model in Table 1, it shows that the bite measure has a negative sign for both RE model specifications. This was expected, as it shows that a higher treatment intensity has a negative effect on the industry's profit margin. However, this negative effect is not statistically significant at any common significance level. Surprisingly, the estimation of the main interest in this study, the DiD coefficient, identifies two interesting facts. On the one hand, this coefficient has a positive sign, which implies that highly affected industries profit margins increased after the minimum wage introduction. This is barely in line with the former study of Bossler et al. (2018), which finds a significant decline in firms' profits due to the German minimum wage. On the other hand, only the FE with TFE specification gives a significant effect of the interaction term's coefficient. Despite this, Table 1 shows the significant level is only at 10%. As the results for the interaction term are either not significant or only at the 10% level, drawing final conclusions from these results might be too shortsighted. Furthermore, all coefficients in the continuous model are of similar size, and thus, the results do not highly depend on the model specification.

In addition to the results of Table 1, looking at Table 2 gives a similar tendency. Similar to the continuous model, the treatment intensity's coefficient goes along with a negative sign. Furthermore, the time dummy for 2015 is for the FE and RE model negative and is in line with the results from the continuous model. Nevertheless, none of the two previously discussed estimates identifies a statistically significant effect on industries profit margins. The DiD coefficient neither reveals a significant decrease of profit margins in any of the four binary model specifications nor is it of a negative sign. Thus, the general results from the binary model are in line with the results in Table 1. However, it is remarkable that every estimation coefficient and standard error in the binary model is smaller than in the continuous model.

To sum up the results of the Continuous and the Binary model, either no strong effect or no effect on firms' profit margins, as a response to the minimum wage, can be identified. This result is robust over all model specification between these two models.

Additionally, the appendix includes further specifications of equations (1) and (2), namely the same analysis is performed for a log level model and a growth rate model. This is done to account for a possible dependence of the final results on the specific model used in the analysis. In line with the results of Tables 1 and 2, the log level model and the growth rate model, fail to identify a significant decrease in firms profit margin as response to the analyzed policy reform.[6]

8.2. Potential threads and limitations

As emphasized in section 8.1, a conclusion claiming that the German minimum wage had no effects on firms generally, and no effect on firms profits specifically, might be highly misleading. At first, the general effect on firms due to the German minimum wage was not analyzed in this study as its focus was on firms' profit margins. Additionally, this paper does only not identify any significant effects due to multiple possibilities such as through the limitation of the data used due to its aggregation and the firms it includes. However, this should not be seen as a synonym for a nonexistence of any effect.

To not identify any significant effect, whether increase or decrease, might be a result of the data used for the analysis. As discussed in section 5, the German Central Bank's SSP dataset is very limited. A main concern is the aggregation of the data, as the SSP data set gives only the results on the industry level; thus, a firm level analysis that takes the industries into account is not possible. Another limitation stems from the firms included in the SSP dataset, which are mostly large companies than Small and Medium Enterprises. However, especially small and medium enterprises are higher affected by the labor market reform in 2015 than large caps. To specify, an analysis dealing with the interaction of firms pricing, profitability, and productivity could reveal more realistic and robust results. Mayneris et al. (2018) previously applied such an analysis to Chinese firms and explained the nonexistence of a significant decrease in profitability with an increase in productivity. Nevertheless, a significant increase in productivity, which can explain the results for the profit margins, is rather unlikely for the German case. Compared to China, the German labor market is fully developed, and additional productivity gains are rather hard to obtain due to the high level of the status quo level. Additionally, former researchers that dealt with the minimum wage reform in the UK (Draca et al., 2011) identified a significant decrease in profitability. Intuitively, the UK's economy seems to be of a higher similarity to the German economy than China's.

Furthermore, the results in 8.1 do not provide information about firms' profits, but it does give information about firms' profit margins defined as the EBT to revenue ratio. Therefore, a decrease in a firm's pre-tax earnings does not necessarily mean a decreasing EBT to revenue ratio. In the German case, Bruttel (2019) found a significant decrease in firm's profits; however, they defined profit as an absolute number of the difference between revenue and the main costs and do not consider profitability. Nevertheless, a decrease in

[6] Note that the same two definitions of the treatment intensity apply to the log level model and the growth level model. Therefore, in total 16 (4 binary log level model, 4 continuous log level model, 4 binary growth level model, 4 continuous growth level model) regressions are performed and presented in the xappendix. The above stated results do not change among different models or different model specifications.

Table 1: Profit Margin – Continuous Model

Standard errors in parentheses. $*p < 0.10, **p < 0.05, ***p < 0.010$

	(1) FE	(2) RE	(3) FE with TFE	(4) RE with TFE
Bite	0	-5.745	0	-5.638
	(.)	(5.777)	(.)	(4.317)
Dummy 2015	-0.369	-0.331	0.117	0.220
	(0.278)	(0.288)	(0.434)	(0.499)
DiD	3.025	2.941	3.368*	3.191
	(1.921)	(1.991)	(1.934)	(2.221)
Observations	434	434	434	434

Table 2: Profit Margin – Binary Model

Standard errors in parentheses. $*p < 0.10, **p < 0.05, ***p < 0.010$

	(1) FE	(2) RE	(3) FE with TFE	(4) RE with TFE
Bite	0	-0.886	0	-0.847
	(.)	(1.394)	(.)	(1.073)
Dummy 2015	-0.249	-0.204	0.247	0.362
	(0.241)	(0.253)	(0.415)	(0.479)
DiD	0.607	0.567	0.646	0.565
	(0.444)	(0.464)	(0.445)	(0.513)
Observations	434	434	434	434

absolute profits does not necessarily imply a decreasing profitability.

Generally spoken, firms can adjust their higher labor costs which are associated with the minimum wage through many different adjustment channels. Namely, they can reduce the total hours worked to keep their total labor costs on a similar level. Additionally, firms can cut or reduce certain non-wage benefits, or they invest less in the training of their employees because these trainings are associated with additional costs. Furthermore, one of the most common responses might be the increase in prices for their products and services supplied. Therefore, consumers would primarily pay the introduction of a minimum wage (Hirsch et al., 2015).[7] In previously conducted research regarding the German labor market reform, Bruttel (2019) identified significant price increases in highly affected industries. Though, they cannot show a significant change in labor productivity, other studies (e.g. Bossler & Boszeit 2017, Pusch & Rehm, 2017) find an increase in job satisfaction which could be an indicator for the efficiency wage model. As discussed in section 3, a positive effect on firms' profits can exist in the context of the efficiency wage model.

All these other possible adjustment channels, combined with the laid-out drawbacks of the dataset that was used, might be a potential factor for the obtained results in section 8.1 of this thesis.

8.3. Placebo regressions

Since the common trend assumption (CTA) is one of the key assumptions to interpreting results of the Difference-in-Difference analysis as causal, this section presents the results of the performed placebo regressions. As already discussed in section 7, the CTA cannot be directly tested; however, placebo regressions test if prior to the minimum wage introduction, a significant 'placebo' effect of the variable of interest took place. Such a significant placebo effect leads to two conclusions: a parallel trend assumption cannot be verified and is unlikely, and the estimated results for the analyzed policy reform cannot be interpreted as causal effect in response to the policy change (Schmitz, 2017). In this study, placebo regressions for the continuous model and the normal model, both with its four different specifications, are performed. In addition, the appendix includes placebo regressions for the four model specifications of both the log level model as well as the growth rate model. Summing up the overall results, the analyses fail to identify significant placebo effects for all specifications in the log level model, the growth rate model, and the 'normal' continuous model. Contradictorily, the two fixed effects specification in the 'normal' binary model reveal

[7] See Harasztosi and Lindner (2019) who conclude that mainly consumers pay for the minimum wage reform in Hungary, instead of firm owner through lower profits. See section 4 for details.

a placebo effect at the 10% level. These results do not entirely support the applicability of the CTA in this context. Thus, the findings shown in section 8 must be interpreted carefully, as the placebo regressions do not allow to fully rely on the CTA. Nonetheless, the graphs shown in section 6 give a slightly different conclusion. The graphs show that since 2010 profitability in exposed and non-exposed industries has not always been on a common trend; however, this was clearly the case for the two years prior the implementation of the minimum wage. Additionally, the control and the treatment group have developed only slightly different for the years 2010-2012. Thus, it would be interesting to apply an adjusted trend assumption to the standard regressions from equation (1) and (2), as previously done in the context of minimum wage research namely by Bruckmeier and Becker (2018) or Bossler et al. (2018), to evaluate if the results would change since the graphical analysis and the placebo regressions over all models do not provide a clear result.

9. Discussion and conclusion

Nearly four million workers have been eligible for the new German minimum wage of € 8.50 per hour. Due to the high ambiguity of the theoretical effects of minimum wages on economic aspects—such as employment, poverty, welfare, firm profits, or consumer's prices—this study addressed firm's profitabilities on the industry level as one major possible adjustment channel of firms' behavior, which is of high relevance for a better understanding of the German minimum wage.

Surprisingly, this study did not identify a significant decline in firms' profit margins in highly affected industries as a response to the minimum wage. Therefore, the results in this thesis are nearly in line with previous international literature on profits (Harasztosi & Lindner, 2019), but still in contrast to Bruttel (2019) review of the effects of the German minimum wage, as he finds a clear decrease in firms' profits. A missing effect in firms' profits would be in line with the theoretical monopsony model from section 3; however, the price increases in Germany (Link, 2019) contradict the applicability of the monopsony model in the context of the German minimum wage. This shows the difficulty in applying a suitable theoretical model and emphasizes the importance of empirical research in the field of minimum wage reform. The results in this thesis are robust over a variation of different model definitions and model specifications. The performed robustness checks reveal two specifications for a significant placebo effect at the 10% level. Combining these results and the graphs shown in section 6, neither a clear verification of the CTA can be made, nor an obvious violation can be detected.

Unfortunately, this thesis faced some limitations according the availability of the data used to clearly apply a Difference-in-Difference approach for the German minimum wage. Therefore, further research can give more enhanced results in the following way: first, new data points such as for the years after 2017 may allow for a more robust and reliable analysis, as currently only the 2 years following the introduction of the minimum wage (2016 and 2017) are included, whereas 5 years of pre-reform data is available. Second, a better understanding of a German firm's response regarding to their profitability can be provided by a less aggregated data set that includes data on the individual firm level, and additionally provides information for results for each region and industry. Furthermore, a smaller degree of aggregation makes it easier to include suitable control variables into the regression, which may change the results. Many previous papers focus on specific adjustment channels, such as prices, employment, or profits; however, due to the high diversity in the results of these papers, a study such as Mayneris et al. (2018) that focuses on the interdependencies of firms' different opportunities to adjust their behavior, such as through prices, profits, and the productivity could give a deeper understanding of how all the different factors are influenced by the German minimum wage. Thus, such a study can also function as a reliable source for policy makers. Additionally, in the near future, an evaluation of whether increases to the German minimum wage (e.g. in 2020 to € 9.34) led to a significant change in firms' profit margins can increase the understanding of firms' adjustments. One reason might be, that the first introduction of an hourly wage floor of € 8.50 in 2015 could be primarily passed through consumers, but at a given—yet unspecified—point, firms and shareholders are paying for further minimum wage increases with lower profits, as no more price increases are tolerated by consumers.

Thus, further research will allow for an interpretation of the missing effects on firms' profit margins in this thesis as it could answer whether there was no effect because the hourly wage floor was not high enough or whether insufficient data led to the underlying results. Indeed, only a joint analysis can reveal the whole story, which means that a conclusion stating the implementation of a minimum wage has no effect on firms' profitabilities, based on these analyses, could be too short-sighted.

References

Ahlfeldt, G. M., Roth, D., & Seidel, T. (2018). The Regional Effects of Germany's National Minimum Wage. *Economic Letters*, *172*, 127–130.

Allegretto, S. A., Dube, A., & Reich, M. (2011). Do Minimum Wages Really Reduce Teen Employment? Accounting for Heterogeneity and Selectivity in State Panel Data. *Industrial Relations: A Journal of Economy and Society*, *50*(2), 205–40.

Bhaskar, V., & To, T. (1999). Minimum Wages for Ronald McDonald Monopsonies: A Theory of Monopsonistic Competition. *Economic Journal*, *109*(455), 190–203.

Bonin, H., Isphording, I., Krause, A., Lichter, A., Pestel, N., & Rinne, U. (2018). Auswirkungen des gesetzlichen Mindestlohns auf Beschäftigung, Arbeitszeit und Arbeitslosigkeit, Studie im Auftrag der Mindestlohnkommission. *Forschungsinstitut zur Zukunft der Arbeit, Evaluation Office Caliendo, Deutsches Institut für Wirtschaft- forschung, Bonn and others*.

Bosch, G. (2015). Der Holprige Weg Zum Gesetzlichen Mindestlohn in Deutschland. *Sozialer Fortschritt*, *64*(7), 173.

Bossler, M., & Gerner, H.-D. (2016). *Employment Effects of the New German Minimum Wage: Evidence from Establishment-level Micro Data., Nuremberg, IAB Discussion Paper; 10/2016.*

Bossler, M., Gürtzgen, N., Lochner, B., Betzl, U., Feist, L., & Wegmann, J. (2018). *Auswirkungen des gesetzlichen Mindestlohns auf Betriebe und Unternehmen. Institut für Arbeitsmarkt- und Berufsforschung, IAB-Forschungsbericht 04/2018, Nuremberg.*

Brown, C., Gilroy, C., & Kohen, A. (1982). The Effect of The Minimum Wage on Employment and Unemployment. *Journal of Economic Literature*, *20*(2), 487–528.

Bruckmeier, K., & Becker, S. (2018). *Auswirkung des Mindestlohns auf die Armuts- gefährdung und die Lage von erwerbstätigen Arbeitslosengeld II-Bezieherinnen und -Beziehern, Studie im Auftrag der Mindestlohnkommission, Institut für Arbeitsmarkt- und Berufsforschung, Nuremberg.* .

Bruttel, O. (2019). The Effects of the New Statutory Minimum Wage in Germany: A First Assessment of the Evidence. *Journal for Labour Market Research*, *53*(1), 1–13.

Bruttel, O., Baumann, A., & Dütsch, M. (2018). The New German Statutory Minimum Wage in Comparative Perspective: Employment Effects and Other Adjustment Channels. . *European Journal of Industrial Relations*, *24*(2), 145–62.

Bundesregierung. (2019). *Mehr Unterstützung für Azubis.* Retrieved from https://www.bundesregierung.de/breg-de/aktuelles/mehr-unterstuetzung-fuer-azubis-1588780[(Accessed 28 Aug. 2019s)

Caliendo, M., Fedorets, A., Preuß, M., Schröder, C., & Wittbrodt, L. (2017). *The Short-Term Distributional Effects of the German Minimum Wage Reform. IDEAS Working Paper Series from RePEc: IDEAS Working Paper Series from RePEc, 2017.*

Caliendo, M., Fedorets, A., & Schröder, C. (2018). Mindestlohn: Stundenlöhne Steigen, Aber Monatsentgelte Stagnieren, . *DIW Wochenbericht*, *85*, 599–608.

Caliendo, M., Schröder, C., & Wittbrodt, L. (2019). The Causal Effects of the Minimum Wage Introduction in Germany: An Overview. *German Economic Review*, *20*(3), 257–292.

Card, D. (1992). Using Regional Variation in Wages to Measure the Effects of the Federal Minimum Wage. *Industrial and Labor Relations Review*, *46*, 22–37.

Card, D., & Krueger, A. B. (1994). Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania. *American Economic Review*, *84*, 772–793.

Card, D., & Krueger, A. B. (1995). *Myth and Measurement: The New Economics of the Minimum Wage*. Princeton, NJ, Princeton University Press.

Card, D., & Krueger, A. B. (2000). Minimum Wages and Employment : A Case Study of the Fast-food Industry in New Jersey and Pennsylvania ; Reply. *The American Economic Review*, *90*(5), 1397.

Destatis. (2008). *Klassifikation der Wirtschaftszweige.* Retrieved from https://www.destatis.de/DE/Methoden/Klassifikationen/Gueter-Wirtschaftsklassifikationen/Downloads/klassifikation-wz-2008-3100100089004.pdf?__blob=publicationFile

Destatis. (2016). *4 Millionen Jobs vom Mindestlohn betroffen.* Retrieved from https://www.destatis.de/DE/Presse/Pressemitteilungen/2016/04/PD16_121_621.html;jsessionid=705175BC44B7086EC1178D19FB3DE548.internet711 (Accessed 28 Aug. 2019)

Destatis. (2019). *Verdienststrukturerhebung.* Retrieved from https://www.destatis.de/DE/Themen/Arbeit/Verdienste/Verdienste-Verdienstunterschiede/Methoden/Erlaeuterungen/erlaeuterung-Verdienststrukturerhebung.html?nn=206824 (Accessed, 28 Aug. 2019)

Draca, M., Machin, S., & Van Reenen, J. (2011). Minimum Wages and Firm Profitability. *American Economic Journal: Applied Economics*, *3*(1), 129–51.

Dube, A., Lester, W. T., & Reich, M. (2010). Minimum Wage Effects Across State Borders: Estimates Using Contiguous Counties. . *The Review of Economics and Statistics*, *92*(4), 945–64.

Fedorets, A., & Filatov, A. (2018). *Great Expectations: Reservation Wages and the Minimum Wage Reform. IDEAS Working Paper Series from RePEc: IDEAS Working Paper Series from RePEc, 2018.*

Garloff, A. (2016). *Side effects of the new German minimum wage on (un-)employment: first evidence from regional data, IAB Discussion Paper 201631, Institute for Employment Research, Nuremberg, Germany.*

Gramlich, E. M. (1976). *Impact of Minimum Wages on Other Wages, Employment, and Family Incomes. Brookings Papers on Economic Activity: BPEA: 409-61. Web.*

Handelsblatt.com. (2019). *Finger weg von einer weiteren Mindestlohnerhöhung.* Retrieved from https://www.handelsblatt.com/meinung/kommentare/kommentar-finger-weg-von-einer-weiteren-mindestlohnerhoehung/23809762.html?ticket=ST-2877840-CydhAernnGta4N43pWGO-ap5 (Accessed 15 Sep 2019)

Harasztosi, P., & Lindner, A. (2019). *Who Pays for the Minimum Wage? American Economic Review 109.8: 2693-727. Web.*

Hirsch, B., Kaufman, B., & Zelenska, T. (2015). *Minimum Wage Channels of Adjustment. Industrial Relations 54.2: 199. Web.*

Holtemöller, O., & Pohle, F. (2017). *Employment Effects of Introducing a Minimum Wage: The Case of Germany. IDEAS Working Paper Series from RePEc: IDEAS Working Paper Series from RePEc, 2017. Web.*

Krasniqi, M. (2007). *Changing Attitudes towards Minimum Wage Debate: How Is The Neoclassical Economic Theory Holding in the Face of a New Era of Minimum Wage Studies? IDEAS Working Paper Series from RePEc: IDEAS Working Paper Series from RePEc, 2007. Web.*

Link, S. (2019). *The Price and Employment Response of Firms to the Introduction of Minimum Wages.*

Mayneris, Poncet, & Zhang. (2018). Improving or Disappearing: Firm-level Adjustments to Minimum Wages in China. *Journal of Development Economics*, *135*, 20–42.

Michl, T. R. (2000). Can Rescheduling Explain the New Jersey Minimum Wage Studies? *Eastern Economic Journal*, *26*(3), 265–76.

Mindestlohnkommission. (2016b). *Erster Bericht zu den Auswirkungen des gesetzlichen Mindestlohns, Bericht der Mindestlohnkommission an die Bundesregierung nach §9 Abs. 4 Mindestlohngesetz. techreport, Mindestlohnkommission.*

Neumark, D., & Wascher, W. (2000). Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania: Comment. *American Economic Review*, *90*(5), 1362–396.

Neumark, D., & Wascher, W. (2008). *Minimum Wages*. Cambridge, MA, MIT Press.

Pusch, T., & Rehm, M. (2017). Positive Effekte des Mindestlohns auf Arbeitsplatzqualität und Arbeitszufriedenheit. *Wirtschaftsdienst*, *97*, 409–414.

Rebitzer, J., & Taylor, L. (1995). The Consequences of Minimum Wage Laws: Some New Theoretical Ideas. *Journal of Public Economics*, *56*(2), 245.

Robinson, J. (1933). *The Economics of Imperfect Competition*. London, Macmillan and Co., Ltd.

Salop, S. C. (1979). Monopolistic Competition with outside Goods. *The Bell Journal of Economics*, *10*(1), 141–56.

Schmitt, J. (2013). Why does the minimum wage have no discernible effect on employment. *Center for Economic and Policy Research*, *22*, 1–28.

Schmitz, S. (2017). *The Effects of Germany's New Minimum Wage on Employ-*

ment and Welfare Dependency. Berlin, 2017. Print. Diskussionsbeiträge / Fachbereich Wirtschaftswissenschaft Economics; 21.

Shapiro, C., & Stiglitz, J. E. (1984). Equilibrium Unemployment as a Worker Discipline Device. *The American Economic Review*, *74*(3), 433–44.

Stigler, G. J. (1946). The Economics of Minimum Wage Legislation. *The American Economic Review*, *36*(3), 358–65.

Stock, J. H., & Watson, M. W. (2015). *Introduction to Econometrics. Updated 3. Ed., Global ed. Boston Munich Print. The Pearson Ser. in Economics.*

Vom Berge, P., & Weber, E. (2017). *Beschäftigungsanpassung nach Mindestlohneinführung: Minijobs wurden teilweise umgewandelt, aber auch zulasten anderer Stellen. (IAB-Kurzbericht, 11/2017), Nuremberg, Germany.*

Welch, F. (1974). Minimum Wage Legislation in the United States. *Economic Inquiry*, *12*(3), 285.

Junior Management Science 6(1) (2021) 39-59

Junior Management Science

journal homepage: www.jums.academy

The Glass Cliff – Women's Thrive to Save Poor Performance and how to Approach it in the Workplace

Julia Vetter

ESCP Europe

Abstract

For decades the rise of women to leadership positions in the workplace has been a conspicuous matter on a global spectrum. Women are challenging the prospects that were once emplaced upon them and continue to break the forefronts of opportunity that face them. This thesis brings life to Eleanor Roosevelt's quote "A woman is like a tea bag: you never know how strong she is until she gets into hot water" by exploring the glass cliff phenomenon. Stating that women have a higher chance of rising into leadership positions during a time of poor company performance, the research conducted for this paper aims to look further into the background of the principles creating this subtle form of discrimination. To discover the root cause of the glass cliff phenomenon, this thesis aims to understand when, how and why these positions are conceived furthering on how to approach the changes coveted by modern society regarding female leadership. This composition takes into account an amalgamation of existing research and the individual empirical research conducted, explaining the causation behind the glass cliff through understanding the biases, stereotypes and societal dynamics that enable glass cliff positions in the workplace.

Keywords: Glass cliff; women in leadership; discrimination; equality; management.

1. Introduction

The rise of women to leadership positions in the workplace has been a quintessential matter on a global spectrum for the last decades. Women have taken on roles in society that do not always fit the perceived image of what it means to be a woman and this courage is what has paved the way for the next woman to do the same. Whether this be women flying airplanes, breaking world records or running Forbes 100 companies; the transparency of gender roles is diminishing. In professional setting as well as in society, women are challenging the parameters that cuffed them to the standards of civilization. There are more women than ever achieving leadership positions although, the majority of companies continue to have male-dominated upper management (Catalyst, 2009; US Bureau of Labor Statistics, 2006). Exploring these new frontiers of leadership and success creates new barriers for many women and minorities climbing the corporate ladder. Barriers that are not always visible yet very restricting in their aptitude.

1.1. Relevance

The tides are turning, socially and institutionally to eliminate the obstacles limiting female leadership. California State Legislature, home to Silicon Valley, passed the Senate Bill 826 (SB 826) on the 30th October 2018, requiring all public companies in California listed on the New York Stock Exchange (NYSE) to have a minimum of one female board member by the closing of 2019. Further prescribing the necessity of women on boards by the closing of 2021 in ratio terms of board members (e.g. minimum of three women directors on a board of six directors or more) (SB 826). Nearing the end of 2018 only 21% of companies with an IPO in California would have met the requirements of the SB 826 according to an Equilar study.[1] This addresses that the implementation of institutional support for women is becoming a requisite in being enforced to create the space for women to access certain positions with less resistance. However, by mandating female leadership, the positions made available could develop into glass cliff positions depending on certain environmental factors that will be further explored in this analysis. Dasgupta and Asgari (2004) state a higher representation of women in positions of organizational leadership may well lead to changes in the associations of leadership and masculinity. Using quotas to regulate the hiring and promo-

[1] Forbes, AllBusiness (2018, October 4).

DOI: https://doi.org/10.5282/jums/v6i1pp39-59

tion processes within working environments causes extensive dispute amongst individuals, justly arguing that talent does not pertain to gender. This thesis will further on the notions regarding quotas in the workplace in respect to their relation to glass cliff positions.

Recent research revealed that economic crisis could present itself as the idealistic time frame for a woman to achieve a leadership position (M. K. Ryan & Haslam, 2005, M. Ryan & Haslam, 2007). One does not have to look far to see glass cliff positions in present day culture; from Marissa Mayer taking over at Yahoo after a long term of crisis within in the company to Theresa May stepping in as Prime Minister after the Brexit referendum in 2016. The leadership roles these women took on had tremendous potential to fail and existed in extreme times of crisis, the very definition of a glass cliff position. Selecting women for these precarious positions has negative implications for their future career opportunities, as directors of unsuccessful companies are less likely to be offered different leadership possibilities (Ferris, Jagannathan, & Pritchard, 2003).

1.2. Academic Foundation

While women are accessing the management positions and breaking glass ceilings around the world, they continue to face a subtle art of gender discrimination along the way (Agars, 2004; King, 2006; Schmitt, Ellemers, & Branscombe, 2003). The significance of gender stereotypes and organizational structures on leadership opportunities for women in the working world is prevalent and well researched by many ambivalent researchers like Michelle Ryan and Alex Haslam. Their studies will be thoroughly exhausted in this thesis providing for the foundation of the research conducted.

Primary research conducted by M. K. Ryan and Haslam (2005) revealed that companies on the London Stock exchange were more likely to hire a woman to their executive board during a time of poor performance in the prior months. They furthered their findings in a laboratory setting three years later, giving participants scenarios of differently performing companies and assessing their attitudes toward future leadership. Men and women partaking in the study showed favoritism towards female leadership during poor performance periods within the company, hence emphasizing that women are preferred in a time of crisis by both men and women (M. Ryan & Haslam, 2007). However, present research also implies that gender discrimination for organizational leadership is not necessarily against women rather regarding the stereotypes about men (Bruckmüller & Branscombe, 2010).

1.3. Research Structure

This analysis aims to understand the causation of these glass cliff in three measures. First, considering the past historical positions women held in the corporate world will further the understanding of *when* the glass cliff emerges. Then, through key analysis of existing literature review and qualitative interviews, this paper will assess *how* and *why* glass cliffs are created. To conclude, evaluating the perceptions of men and women in the workplace in regard to the idealistic working environment, this thesis will shed light on what the ideals in the working world could establish when diminishing gender discrimination.

The research conducted is contingent on the qualitative interviews executed in the time frame provided for this thesis. Through the analysis of the responses collected and further interpretation, the structure this paper takes on derives from the presumptions made by the participants and according research.

Chapter 2: The literature review summarizes existing and essential insights to understand the foundation of the glass cliff phenomenon. By defining the indispensable terminology relevant to deriving the means of the phenomenon, this chapter will lay the ground work needed for the research phase to initiate. This latter part of this chapter will provide a holistic view of the different phenomena associated to the glass cliff and the main attributions that explain its existence.

Chapter 3: The methodological approach depicts the research method used in this thesis. First, by covering the general approach then thoroughly discussing how the qualitative research was designed by capturing the empirical setting, procedure, participants and analysis in depth.

Chapter 4: The results of the interviewees conducted are summarized in this chapter, defining the different insights gathered from the 13 participants. The analysis is split into three sub categories based on the point of time in the interviewees career path: (1) the younger generation, (2) the middle generation and (3) the mature generation. The chapter identifies the primal patterns in the responses collected and compares the view points of the different contributors.

Chapter 5: The discussion of this paper interpretation of the major findings concluded from the interviews conducted. Furthermore, discussing the meaning these findings contributed to the thesis and how the research is relevant in regards to prior existing work. Lastly, this chapter will also analyze the validity of the research conducted.

Chapter 6: The conclusion and contribution will summarize the main findings that were drawn from the research conducted. It will assess the future of female leadership in the workplace and regard the limitations that this research encountered. This chapter will also identify future research that can be conducted to further understand how female leadership is impacted by gender discrimination and how women potentially add to the discrimination.

2. Background

To understand the glass cliff and how it was created, one must first understand the psychological phenomena that underlie its existence. It is critical to recognize the differences between a glass cliff and glass ceiling, which this chapter will elucidate. Essentially, the creation of a glass cliff position is expedient in a plethora of different pathways and usually isn't due to only one of the factor. While stereotypes about

gender and leadership are a main causation, biases, group dynamics and perceptions of change can significantly impact the reasons behind the formation of a glass cliff. This chapter will decipher the main causes that equate to the creation and continuance of this phenomenon.

2.1. Definitions

As the glass cliff phenomenon pertains to other phenomena regarding the path to leadership for women and minorities, this chapter will clarify the difference in the conceptions and their significance parallel to the glass cliff. Understanding the differences between terms such as the glass ceiling and glass elevator is critical for the apperception of the glass cliff, which will be defined in the aftermath of the former.

2.1.1. The Glass Ceiling

The glass ceiling by definition is *"an intangible barrier within a hierarchy that prevents women or minorities from obtaining upper-level positions."*[2] It is argued that women are held back from corporate promotion and advancement due to informal networks of communication and male preconceptions and stereotyping of women (Hefferman, 2002). Therefore, creating a glass ceiling that permits for women to see the success they could have while climbing the latter of success though rarely ever reaching the promotion. While the glass ceiling has hindered women from achieving managerial success in the past, the number of women who occupy leadership positions is greater than ever (Bullard & Wright, 1993; Bureau of Labor Statistics, 2005; Dreher, 2003; Equal Opportunities Commission (EOC), 2002).

2.1.2. The Glass Escalator

While discussing gender phenomena, the glass elevator is often overlooked. The term refers *"to the precipitous promotion of men over women into management positions in female-dominated fields such as nursing, education and social work"*.[3] Men often experience positive discrimination when working in female- dominated fields, where women are eager to work with men in *"their"* occupations, unlike the *"poisoned"* work environment women encounter when entering into male-dominated occupations (Williams, 1992). Discrimination against men in certain female-dominated occupations is still prevalent today, starting in the youthful years when teenage girls start their first babysitting jobs, teenage boys are less likely to be chosen to care for children than girls are. This later pans out into professional careers where the initial discrimination unravels creating female- dominated occupations. However, when men do stay in the female-dominated occupation, they are more likely to reach managerial success for instance, as a principal of school.

[2] Glass Ceiling (n.d.) in Merriam-Webster's collegiate dictionary.
[3] Broner, Asselmeier, and Broner (2013, September 30).

2.1.3. The Glass Cliff

Essentially, the glass cliff phenomenon states that *"women's perceived suitability for senior positions is likely to increase under conditions of organizational crisis."*[4] The term *glass cliff* was coined by S. Alexander Haslam and Michelle K. Ryan (Haslam & Ryan, 2008; M. K. Ryan & Haslam, 2005) as a rebuttal to Judge (2003) analysis of the highest Financial Times Stock Exchange 100 (FTSE 100) performing companies' boards. He observed that in the ten highest ranked companies with women on their boards, six had underperformed the mean performance of FTSE 100 companies. Meanwhile, the five companies with the lowest amount of female board members performed better than the FTSE 100 average. His comment that "corporate Britain may well be better off without women on the board"[5] was what triggered Haslam and Ryan to analyze the correlation of company performance and appointments of male or female board members (M. K. Ryan & Haslam, 2005).

2.2. Causes of the glass cliff

To understand how, why and when glass cliffs are created, one must look past the current corporate positions and return to the origins of modern societies perceptions. Due to a large part of historical leadership having been male- dominated, stereotypes regarding the perceived qualities of gender and leadership shaped the biases of women in leadership positions. This continues influencing the perceptions regarding women in general and the qualities of success, creating stereotypes that strengthen the feasibility of creating glass cliffs in the workplace. Further theories, biases and stereotypes that lead to the creation of glass cliffs will be explored in this chapter.

2.2.1. Implicit theories

There is bountiful research that affirms a correlation of gender differentiation in the workplace, more specifically a reflection in people's implicit theories about gender and leadership. Evidence observed that there is an incompatibility in perception of beliefs about what it means to be female and what it means to be a good leader. (Agars, 2004; Eagly & Karaus, 2002; Heilman, Block, Martell, & Simon, 1989; Schein, 1973, Schein, 1975, Schein, 2001). Implicit theories about management and gender are critical as they are not only descriptive but also powerfully descriptive (Heilman, 2001) meaning that they have immense impact on how women are treated in the workplace as these theories provide the foundation for expectations of how roles are distributed and how to act in the workplace.

2.2.2. Stereotypes about Gender and leadership

While it is often underestimated, stereotypes are a main contributor to differences in associations of leadership styles. Stereotypes are *"something conforming to a fixed or general pattern"*[6] (See Appendix 1). Through societies prejudiced

[4] *Academy of Management Review* 2007, Vol 32, No.2, p. 554
[5] Judge (2003, p. 21)
[6] Stereotype (n.d.) in Merriam-Webster's collegiate dictionary.

and overgeneralized view of gender roles and leadership qualities, the expediency of creating a glass cliff is accelerated. In a historical context, leadership has been predominately male which has subconsciously shaped societies' view of how leadership is portrayed, creating a masculine stereotype of leadership (Dasgupta & Asgari, 2004; Eagly, 2007). The effect that gender stereotypes have on women in the workplace have been extensively researched and have been expanded into two phenomena, think manager- think male and think crisis- think female, which will be further expanded on in this thesis. Having studied women in management to ascertain their issues and challenges in managerial diversity, Antal and Izraeli (1993) stated that *"probably the single most important hurdle for Women in management... is the persistent stereotype that associates management with being male."*[7]

Think manager—think male

One of the major hurdles women encounter when climbing the steps of leadership is the perception of what manager qualities are, quite in contrary as to how women are; this is more simply reduced to the phenomenon of *think manager-think male* (TMTM). Schein (1973) studied the correlation of stereotypes of men and women as well as middle managers. Participants were given a list with ninety-two descriptive terms and were asked to associate them with either (1) women in general (2) men in general or (3) successful middle managers. She observed that sixty of the terms overlapped between the characteristics of men and managers including terms such as self-reliant, confident, ambitious, forceful and informed. Schein repeated the study with male (1973) and with female (1975) managers, concluding the same results in both studies. Schein's study was repeated on a global scale in the United States, United Kingdom, Germany, China and Japan as well as in the Military[8] all proving the initial findings of Schein (1973). Schein's study proved that there is direct linkage between the perceptions of women and the perceptions of what it means to be a good manager. The TMTM phenomenon is seen even stronger in management areas that have been male- dominated where women are then at a twofold disadvantage against a male candidate. (Schein, 2001). Furthermore, the TMTM is evident in the mannerisms that managers portray during the workday; if a man is assertive then he's being a leader whereas, if a woman is being assertive she is seen as pushy (Eagly, Makhijani, & Klonsky, 1992) Therefore, the TMTM phenomenon is a lose- lose situation for women that continues to be an obstacle in the workplace and is a main contributor to the creation of the glass cliff.

Think crisis—think female

Context determines the grounds for any preconceptions to exist hence why the prejudice towards female leadership changes when there is crisis involved.

Known as the *think crisis- think female* (TCTF) phenomenon (M. Ryan & Haslam, 2007), the correlation that women may be better equipped in situations of plight. Schein (1973), Schein (1975) study also revealed that the descriptive terms associated to women were understanding, intuitive, helpful, sophisticated and cheerful; traits that are often desired and useful during a time of crisis. TCTF was further confirmed when a three-part study (M. K. Ryan, Haslam, Hersby, & Bongiorno, 2011) compared gender stereotypes in successful and unsuccessful companies resulting in apparent correlation between the preference for male leadership in successful companies and female leadership in the unsuccessful companies. Table 1 is a result of Study 1 and reflects current perception of characteristics of male and female managers in successful and unsuccessful companies. The TMTM association is evident in the results, more specifically underlining that when one thinks of crisis they think *"not male."*

However, Table 2 is the result of Study 2 which underlines the qualities desired from managers in successful and unsuccessful companies, the mere difference in length in the qualities exhibited between men and women reflects the stereotypes attributed to crisis management and women. It also shed light on the ideology that female qualities are much more desired in successful companies than currently expressed due to the status quo bias of men having previously dominated upper management.

Furthermore, as seen in Table 1 the TCTF association is evident when the company was performing poorly and needed a responsible, managing leader that could endure the hardship. However, the TMTM association is visible in the more agentic roles like spokesperson and improvement of company standing where more masculine traits were desired.

TCTF is part of the foundation that creates glass cliffs, it puts women in the position to help in a crisis but exposes them to greater risk by doing so. By inheriting prior poor management, the woman will encounter heavier criticism and be judged more intensively than taking over during a time of success within the company. It's easier to make her be the *"fall guy"* than a man during critical times as the prejudgments regarding her management skills are already in place.

2.2.3. Status quo bias

Why fix a wheel if it's not broken? Peoples reluctance to change something if there is no need for it or continuing to stick to a decision made previously, is known as the status quo bias (Samuelson & Zeckhauser, 1988). Due to a large majority of history having been male-dominated, the status quo bias is applicable to existence of the glass cliff or more specifically the glass ceiling. By having no obligation to change the leadership styles when a company is performing well, men continue in leadership but when poor performance calls for organizational restructuring women enter the solution; this is how the status quo bias amplifies the existence of the glass cliff.

[7] Antal and Izraeli (1993, p. 63)
[8] M. Ryan & Haslam, 2007, p. 550

Table 1: Attributes Seen to Be Characteristic of Mangers of Successful or Unsuccessful Companies and Also Stereotypical of Men or Women (M. K. Ryan et al., 2011)

Attribute	Stereotypical of men	Stereotypical of women
Characteristic of managers of successful companies	Decisive (4.11. 3.98) Dominant (4.11. 4.24) Feeling not easily hurt (3.97. 3.46) Able to separate feeling from idea (3.97. 3.46) Speedy recovery from emotional disturbance (3.94. 3.85) Forceful (3.81. 3.88) Hides emotion (3.69. 3.80) High need for autonomy (3.69. 3.80) Aggressive (3.63.4.10)	Intuitive (3.89. 4.04) Sophisticated (3.89. 3.98) Tactful (3.83. 3.73) Neat (3.80. 4.22) Understanding (3.69. 4.49)
Characteristic of managers of unsuccessful companies	Hasty (3.57.3.41) Quarrelsome (3.43. 3.50)	Fearful (4.24. 3.88) Wavering in decision (4.23. 3.64) Nervous (3.91.3.51) Passive (3.80.3.46) Sympathetic (3.73.4.67) Desire to avoid controversy (3.62. 3.55) Frivolous (3.49.3.38) Understanding (3.46.4.49)

Note: Figure in parentheses refer to the mean level of characterizability for managers and for gender, respectively.

Table 2: Attributes Seen to Be Desirable for Mangers of Successful or Unsuccessful Companies and Also Stereotypical of Men or Women (M. K. Ryan et al., 2011)

Attribute	Stereotypical of men	Stereotypical of women
Desirable for managers of successful companies	Decisive (4.69. 3.98) Speedy recovery from emotional disturbance (4.57. 3.85) Feeling not easily hurt (4.51. 3.65) Assertive (4.33. 4.10) Not uncomfortable about being aggressive (3.57. 3.61) High need for autonomy (3.51. 3.80) Dominant (3.44. 4.24) Hides emotion (3.27. 3.90)	Understanding (4.51. 4.59) Aware of the feeling of other (4.37. 4.30) Tactful (4.33. 3.73) Intuitive (4.27. 4.04) Grateful (4.22. 3.86) Neat (4.10. 4.22) Modest (4.08. 3.79) Sophisticated (3.92. 3.98) Sympathetic (3.63. 4.67)
Desirable for managers of unsuccessful companies	Decisive (4.47.3.98) Assertive (4.30. 4.10) Able to separate feelings from ideas (4.23.3.46) Speedy recovery from emotional disturbance (4.20.3.85) Feelings not easily hurt (3.98. 3.65) Not uncomfortable about being aggressive (3.53.3.61)	Understanding (4.75. 4.49) Intuitive (4.60. 4.04) Tactful (4.60.3.73) Aware of the feeling of others (4.55. 4.37) Grateful (4.38.3.86) Modest (4.18.3.79) Sympathetic (4.15.4.67) Neat (4.15.4.22) Sophisticated (3.85. 3.98) Sentimental (3.47.4.20) Desire to avoid controversy (3.40. 3.55)

Note: Figure in parentheses refer to the mean level of desirability for managers and characterizability for gender, respectively.

2.2.4. Hostile and Benevolent Sexism

Sexism is the mere discrimination based on a person's sex or gender.[9] Ambivalent sexism consists of two frameworks, hostile and benevolent sexism. Hostile sexism has a negative evaluation of gender and suggests that in the workplace the appointment of women to leadership positions is with the mere intention to disadvantage the woman and desire to see them fail (M. Ryan & Haslam, 2007). This is a major factor of influence for women and other minorities in the workplace. While many companies have policies and regulations against discriminatory behavior within the corporation, it doesn't eradicate the individual prejudices certain management positions may hold. Benevolent sexism appears to be more positive yet discriminates against a person in a subtle mannerism. It refers to women taking on or being rewarded with leadership positions as an attractive challenge whereas

[9] Sexism (n.d.) in Merriam-Webster's collegiate dictionary.

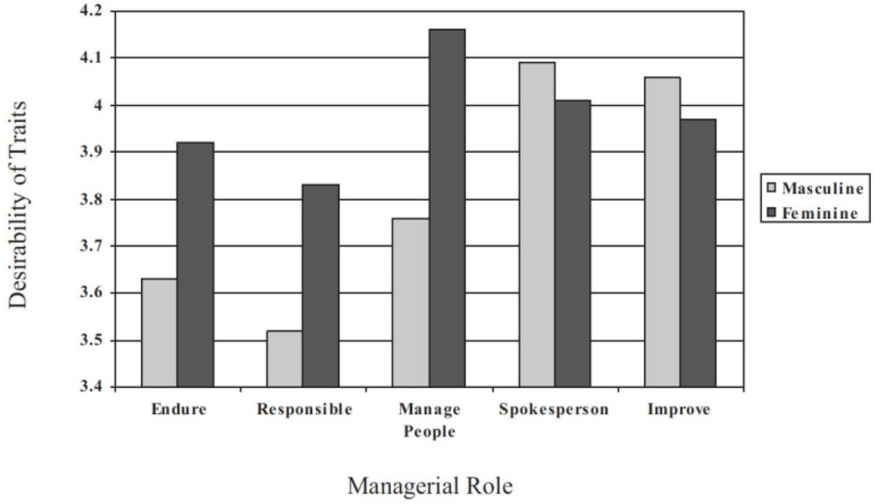


Figure 1: The impact of managerial role on the desirability of masculine and feminine traits (M. K. Ryan et al., 2011)

the position is actually problematic. (Cuddy, Fiske, & Glick, 2004; Glick & Fiske, 2001). Since the advancement is seen as a favor to the woman, it's difficult for her to reject the offer and provides the means for this overt discrimination to occur. These promotions allow for sexists to appear as a supporter of women although they are subtly maintaining the gender hierarchy within the company by ensuring that the professional development of the women promoted remains limited. Thus, any type of sexism affects the implication of the glass cliff for women rising into management positions.

2.2.5. Group Dynamics and In-group Favoritism

Group dynamics shape how an individual thinks and acts which therefore significantly affects the decisions and actions of that person. Though it has been proven that when many share the ideas of another, a bias within the group is created leading to a stronger and more effective ideology. The infamous quote "it is a man's job" is the literal implication of what in-group favoritism is. If a group decides that a certain role can only be handled and executed by a certain figure, for instance that a job can only be done by a man then the bias affects the entire group in favor of that opinion and more specifically any women who could potentially fill that role. This expands to identifying that women do not have the *"homo-sociable"* (Ramsay & Parker, 1992) network that provides them with the support, infrastructure and resources that abet men ascending the corporate ladder and climb the ranks of senior management (Higgins & Kram, 2001; Ibarra, 1993; Ragins, Cotton, & Miller, 2000; Tharenou, Latimer, & Conroy, 1994). Due to the favoritism within the group, during the decision making of leadership advancement this dynamic easily restrains the opportunities for women to advance in management, therefore directly aiding to the creation of a glass cliff.

2.2.6. Signaling change

Appointing a woman into a precarious leadership position after having only men in the prior symbolizes a change in the situation. The purposes behind the sudden leadership change can be one of many, signaling a change to internal and external observers. New company policy to encouraging gender equality can encourage the companies' willingness to apprehend its prior methods of managing thereby alleviating the TMTM principles within the corporate culture. Signaling change can also be seen in a more sinister perspective, akin to a company having nothing left to lose. This can be seen in Kaplan and Minton (1994) study in Japan where the appointment of a highly visible *"outsider"* (e.g. a woman) is associated with poor company performance, therefore influencing the impression of the management. With changes in leadership signaling such perceptions inside and outside the company, the stigma created directly induces the creation of a glass cliff.

3. Methodology

The focus of this chapter is to explain the research strategy and general approach of this thesis. Through a combination of theory and practice, this paper will focus on an academic analysis of the current perceptions in the workplace regarding women in leadership. After the literature review that set a foundation for the research to proceed, the qualitative research conducted will provide diverse insights into today's working environment. The research will be explained in depth through the empirical setting, target group, procedure and data analysis.

3.1. General Approach

After assessing the amount of research already existing, in the genesis of this thesis it was clear that an empirical approach is justified. This qualitative research has a theoreti-

cal base that will interpret context through sensing and exploration of relevant factors. Through inductive reasoning, this small sampled, non- standardized approach will provide a contextual understanding. Through a combination of exploratory and action research, the participants involved will be asked to reflect on their personal experiences subject to analysis in their actions and speculations.

3.2. Qualitative Research

Qualitative research is beneficial in understanding the conceptual orientation of human beings and their surroundings. Having an explorative character, using a qualitative approach to analyze how women are working their way up the corporate ladder seems only fit. This thesis intends to capture not only the women's perspective but to embody the different points of view that contribute to the leadership appointments today.

3.2.1. Research Design/ Empirical Setting

To ensure that this empirical research is well rounded and reflects a varied origin of opinions, the empirical setting is eminently diversified. This holistic approach drove the participants to be from a plethora of countries in assorted industries. This includes the medical and educational field, the entrepreneurial world and consulting, real-estate and racecar industries. Furthermore, the research conducted was on a global scale collecting insights from participants in Italy, the United Kingdom, the Netherlands, Germany and the United States. By providing a scope of backgrounds, including male and female dominated industries, this thesis will provide a perceptive insight across the different paths to leadership.

3.2.2. Target group/Participants

To begin a research study, one must first derive the target group. The path to leadership will be best reflected by the women who have already embarked on journey. Among the women climbing into leadership, the women in upper management positions are the initial subjects of this thesis making them a critical component to shape the focal point of the research conducted. By having the women in management positions reflect the dependent factor it allows permits their age, location and industry to be the independent factors. By fixating this target group as the centrifuge of this research, further participants can be added to provide a more perceptive insight.

For a broader understanding of the women and the different industries, it was decided to add a few outliers to the portfolio to allow for a more contextual analysis and limit any feminist biases. This drove to including younger and older participants ranging from twenty to eighty-seven, both male and female in different statuses of life. By including a scope of age ranges, a distinction can be made in the changes that have occurred with time. This permits for a comparison to made between the younger perception joining the workforce and looking into the future for their careers in regards to a retired point of view that reflects how the workplace has changed and how the present situation was shaped. Furthermore, by assessing both genders point of view, the research provided gives proper insights into the different perspectives of how leadership can be attained within a company; more specifically, it provides the information needed to assess the difference in paths based on gender. By adding outliers, there was the potential of excess information and disagreeing perspectives which is accounted for accordingly in the analysis. The representation of these participants was done through convenience sampling based on availability and willingness to contribute.

3.2.3. Empirical Procedure

The preeminent empirical procedure driving the research concluded was conducting qualitative interviews. Through a combination of personal meetings, phone interviews and written responses, data was collected over a six-week time frame in the spring of concoction of this thesis. To find patterns in the data, there were three different question sets based on the classification of the participant being interviewed, (1) woman in leadership, (2) man in leadership and (3) outsider point of view. By differentiating the content and context of each question set, the focal point of the interviews was maintained. The question set for women included one additional question in regard to the personal reflection of the target group of this research. The women in leadership were guided to give reflective insights on their personal experiences while climbing the corporate ladders in their given industries. This provided easy identification of patterns experienced in across industries in regards to difference in genders when ascending to leadership positions. The men in leadership were conveyed similar questions in regard to their female colleagues being promoted as well the personal instances they had experienced. The male perspective allowed for subjective insight on how men reflected on the concept of female leadership, especially on the discrimination of it. Lastly, the outsider group was asked the same question set as the men in leadership group, this accounted for a more objective point of view into the corporate management world. The question set for men in leadership and the objective point of view had one ancillary question to the female interviews as it asked for the male path to leadership and difference in the perceived path for a woman.

There was only a primal interview conducted with each participant which covered all the themes within the scope of research. Additionally, to ensure a diverse portfolio of participants, demographic analysis followed each interview. The demographic data collected reflects the gender, age, level of education, occupation and nationality. By collecting the demographics of each participant, the anonymity of each interviewee will be kept for the purpose of this thesis with mere reference to their background.

3.2.4. Data Analysis

The resolutions and observations of the participants are the foundation for the empirical analysis conducted. Based on the first-hand experiences of women who have embarked

on the leadership path combined with the views of bystanders who recognized their journey and reflected their contributions, the analysis conducted in this thesis will provide the grounds for interpretation of the current workforce environment in concerning female leadership. Taking into account the different obstacles, networks, industries, characters and demographics, the analysis conducted will apprehend the overall effect of the glass ceiling in the modern corporate world. Based on the answers provided by the participants, specifically the women in leadership interviewed, this thesis will analyze the commentary carefully while elucidating the effects of the glass cliff facing women rising in leadership. Essentially, the qualitative analysis of the data collected will encompass the research question respecting any margin of error and further research to be conducted.

4. Results

This chapter focuses on the insights collected from the interviews conducted, summarizing the different points of views shared in opinions and experiences regarding women's journeys to leadership. To begin, an overview into the research will be provided to administer the proper foundation to analyze the responses accordingly. Following with an introduction of the individuals interviewed giving minor insight into their background. This will be furthered into reflecting on the responses of the interviewees to assess any patterns in the answers collected. Summarizing at the end with a small synopsis on the interviews conducted.

4.1. General

The term lead means *"to direct the operations, activity, or performance of"*[10] which is can be regarded as easier to do with more time or experience. It would be an over generalization to conclude that good leaders need immense experience however it is arguable to say that leading becomes easier with experience and knowledge.[11] The term leadership is derived from lead and defines as the *"capacity to lead."*[12] Therefore, one could imply that leadership is often possessed by individuals that have the right experience to lead which inherently comes with times. Naturally, this also effects the stages when leadership is acquired within the working world; justifying why the average age of CEO's is 57.[13] This factor is critical to note for this paper, as the interviewed individuals come from different age backgrounds and experiences which is evident in the answers collected.

4.2. Meet the interviewees

Consisting of 13 individuals who agreed to partake and contribute to the research conducted, the interviews provided a plethora of different insights into the experiences and perceptions in the modern working world regarding women in leadership (see Table 3).

[10] Lead (n.d.) in Merriam-Webster's collegiate dictionary.
[11] Center for Creative Leadership, n.d..
[12] Leadership (n.d.) in Merriam-Webster's collegiate dictionary.
[13] Fortune (n.p.), n.d..

4.3. General overview in results

Upon conducting the interviews, it was visible to see that opening the conversation about female leadership can be strenuous some individuals as it compelled them to look at oneself and at the people around them to see how they were experiencing their career path. The individuals that were aware of their own actions, or in actions, could to deliberate one's standpoint and perspective on the matter, which appeared to strenuous for some of the participants. This impression was coherent in some of the interviews conducted, most noticeable in the different standpoints in career also reflected in the age range and gender of the interviewee. This chapter will highlight the results regarding women in leadership positions from different standpoints which have been divided into three sub- patterns (a) the younger generation, (b) the middle generation and (c) the mature generation.

4.3.1. The younger generation

The younger generation, age range 20-29, was eager to discuss and challenge the points of view at hand. The main point of differentiation with this group was that the majority of the participants still have their careers in their future, merely anticipating the working world ahead of them based on their own first experiences and the impacts left on them by their forerunners. This age group showed the least bias and heavily emphasized their wishes for truly equalitarian opportunity when ascending the ladder of success while still upholding meritocracy. One could speculate and suggest that they were sanguine for their own futures hoping that the past and current hurdles and limitations might cease by the time they embark on their career paths. This generational difference depicted some of the more passionate points of view on the topic during the discussions. The women heavily emphasized their own role in society to change the situations at hand.

This age faction reflected the futuristic view imposed on their generation, the insights these individuals expressed showed a pattern that each of them felt they still had to prove themselves as women in comparison to men. When asked to further on their angst regarding their future career paths, all five women expressed that they knew that pure performance, merit and hard-work was not going to suffice for their promotions. Interviewee C stated *"men are expected to be the store managers, the district managers and the CEO's... women have to work a lot harder to get [into leadership positions] those, and more importantly feel worthy and respected in that position."* She continued to express that the stereotype associated with female CEO's continues to restrict the influence and ability a woman can have in the workplace, arguing that Hollywood's image of a female CEO has painted them into *"sassy, sexy and smart, with a tight pant suits, walking into work in the court heels and their skinny chai latte."* While her argument maybe premature, there is veracity in her statement; how Hollywood has portrayed female leaders in movies is a reflection of how society views women managers. From Miranda Priestly in the Devil Wears Prada to Claire Underwood in House of Cards, the view on negative

Table 3: List of Respondents (Data compiled by Author)

Interviewee	Gender	Age	Origin	Education	Job Title
A	Female	20	United Kingdom	Student in medical school	N/A
B	Female	21	United Kingdom	UK GCSE-A Level	Lettings negotiator/administrator
C	Female	21	United Kingdom	Undergraduate Degree	Student of midwifery
D	Female	27	United Kingdom	Undergraduate Degree	SaaS Sales
E	Female	27	United Kingdom	Postgraduate Degree	Teacher
F	Male	30	Germany	Engineering	Consultant
G	Male	34	United Kingdom	UK GSCE - Level 2 & 3	Director
H	Female	37	Italy	MSc. Language/Literature	Fee Lance/Race car
I	Male	38	United Kingdom	Law Degree	Life Coach
J	Male	47	United Kingdom	Bachelors Degree	Entrepreneur (PMO/Coaching/Event Mgmt.)
K	Female	54	United State of Ame	Msc. in Finance	Finance/now Consultant
L	Female	56	United Kingdom	Postgraduate Degree	3D Architectural Visualisation
M	Female	57	Netherlands	College	Mastermind Facilitator

female leadership is strengthened as seen in Schein (1973) study of stereotypes about women in leadership. Interviewee A remarked that *"men believe they are superior without knowing their [women's] qualities first,"* reinforcing the argument made by Interviewee E. As a medical student, she reflected that she does not feel at a disadvantage in comparison to men in her field except for certain medical specializations such as orthopedic surgery where major physical strength is required. However, she did comment that the minor activities such as planning, project leads and minor leadership were often a struggle for her and other women in to take lead in. One could see her reflecting on the comment she just made, which led to Interviewee E asking herself what she does about it and what she should do about it.

Interviewee D initiated the conversation of maternity leave upon being asked if she ever found it difficult to advance into leadership as woman. She expressed her concern that once *"the engagement age"* comes around, women are assumed to be having a baby soon and are therefore overlooked for the future promotions. As a 27-year-old, she disclosed that she feels has to *"prove herself even more"* continuing on to say that many of her female friends find it particularly more intense in their late twenties and early thirties to be promoted. The topic of women stepping out of the office to start their families will continue to be a conversation that women, men, companies, governments and societies as a whole need to have. While certain companies have great policies and governments are encouraging parental leave for both women and men, it does not resolve the prejudice affiliated with maternal or paternal leave. The conversation of how the future will look in regards to paternal leave provides space for an entirely new thesis therefore, the conversation with Interviewee D was subdued back to the current state of women in leadership positions.

The interviews conducted also addressed the associated phenomena surrounding the class cliff and Interviewee E spoke about her glass escalator experience. As a school teacher, she works primarily with women coworkers and leads her faculty; however, she stated that the *"senior leadership is very male dominated and there is explicit sexism."* Furthering into the conversation, she said that *"many men seem to have been prematurely promoted [in my organization]. Their confidence and arrogance is taken at face value while their soft skills are lacking in comparison to female colleagues."* She continued to give the example of the glass cliff in her organization by expanding on her experience of joining the school. The school had been receiving negative press due to a scandal and they had just hired a new principal, a woman, after two years of clearing the name of the school, the principal was asked to leave for a back-office job in the district office. She was replaced by a male principal; Interviewee E was unsure what the specifics behind the *"promotion"* were yet showed slight hesitation and confusion when being asked to further on the subject. Examples like the following disclose how subtle the glass cliff positions can occur and disappear; also shining direct light into the presumptions women have on the scenarios.

An interesting pattern arose in this age faction, when being asked about their ideal workplace environments, all four used the word *nurture* in their description. Plus, three of the interviewees used *supportive* in the environmental aspects. These two terms were characterized as female descriptive terms (Schein, 1973) and were reflected to the similar qualities desired of women managers in any company regardless of performance (M. K. Ryan et al., 2011) depicting the terms *understanding* and *aware of the feelings of others*. This strengthens the argument that the qualities making a good manager are not explicitly *"male traits"* and that leadership positions in modern society are yearning to be filled with qualities that currently being associated with the stereotypes of women. After initiating the conversation of glass cliffs in the workplace, all four interviewees shared how they had seen these positions arise for themselves or for the people in their environments. The explicit example from the teacher, Interviewee E, showed that glass cliff positions do not necessarily exist in only male dominated environments but can be created in female- dominated workplaces as well.

4.3.2. The middle generation

"Some people are like trees; they take forever to grow up." This was the first sentence that Interviewee I said before initiating the conversation, immediately setting a different tune for our interview. It quickly made me realize that this stage of interviewees would provide a very different point of view on the matter at hand, a more seasoned and pragmatic outlook. The 30 to 45-year-old perspective did indeed give a sensible insight into the working world of today, considering they are the current middle and upper management climbers working through the exact theme of this thesis. Overall, the conversations proved to be more distant and emotional than with the prior group, the interviewees displayed a sense discomfort in the topic and had very different acumen of how the glass cliff exists and should be addressed in the workplace.

Interviewee I was the first contributor to this age aggregation and was quick to express his frustration towards the conversation as whole. Baffled that in *"modern times, we are still having a conversation about gender and its impact on somebody to do a job"* he clarified that he is an outspoken egalitarian that does not see how gender, skin color, sexuality or religion could impact one's ability to lead. Heavily emphasizing that one's skill, mindset and execution are the mere principles needed to make a good leader in the workplace. Interviewee F also addressed the issue in stating *"the only determining factors that leadership should be measured on are hard-work and merit."* Both were irritated when discussing how female leadership is viewed in this present day, giving examples of inspiring female bosses in their surroundings. As men, they had a distance to their own point of view, showing a sense of empathy towards women and an awareness for the men that hinder the furthering of certain women to leadership positions.

The three men who responded in this age group degraded their own gender in the conversations, they were harsh in their judgements towards their male colleagues and bosses. Not directly blaming themselves for the struggles women have encountered in the workplace, Interviewee F, Interviewee G and Interviewee I showed a sense of acrimony and resentment in the situation. All three men knew women in their inner circles who had faced some form of gender discrimination in the workplace; they shared how they were appalled that there were still men in today's society that could see a woman as less competent for a job. Interviewee F appended that there were still physical differences between men and women and that this can make certain jobs difficult for certain genders. Giving the example of heavy lifting jobs for men that could be difficult for the average woman to do in her *"normal"* strength, vice versa noting that it could be difficult for a man to work in areas where delicate materials are in place or certain sensitivity is required. He discussed that the difference in genders should not be perceived as a negative trait to either one but as a positive value added to each other's lifestyle, stating that *"the best teams I've worked on had an equal balance of men and women that were both open to working towards the best solution together, knowing that every individual had a different value to bring to the table."* He got frustrated when he reflected on his time as an engineer where he was surrounded mostly by men, who then also made every woman in the field feel alone. Upon asking where the frustration came from, he noted that abilities in IT, engineering or finance were not based on gender but merely based on education and that this misconception of *"female"* and *"male"* jobs was one of societies biggest misconstructions.

In an attempt to understand why Interviewee F and Interviewee I were so frustrated with the discrimination defined, the discussion turned to understanding the male side of gender discrimination in the workplace. Interviewee I quickly stated that he believed *"it boils down to insecurities"* amongst men, not stepping into their masculinity to identify the added value women bring to a table. He argued that a real man would never stand in the way of a woman, he would stand behind her to push her forwards; he noted that it's easier when the woman is someone one cares about than just a colleague. *"When losing a promotion to a woman, the man should feel no less than if he lost the promotion to man, however they often do"* shared Interviewee F, he said that it was often the subtlest of actions that showed how biased one can be when it comes to women working their way up the corporate ladder.

Upon discussing quotas, the men shared the view that quotas were necessary in certain industries where women have not yet made the imprint they need to and that it opens the door of opportunity for some. However, they quickly added that it does not resolve the social image of the woman in her leadership position, the sentiment towards the woman promoted merely degrades as the surrounding colleagues view of the promotion as a regulation instead of her merit. This could effectively be worse as the male colleagues may resent the woman or not respect her new position, despite her true ability to do the job. Stating that *"the benefits of quotas make foul play limited and provide the space for the women to rewarded for their skills where they previously were not"* Interviewee I insisted that for the time being, quotas may be the only feasible option to permit for women leadership in certain industries. He also clearly expressed that he wishes the need for quotas and glass cliffs would dissipate in the near future where all leadership positions are filled on the mere *"skill, mindset and execution"* one presents in the workplace.

Incorporating the conversation about glass cliffs, all three interviewees noted that they had seen a woman get promoted during a *"riskier time frame"* in either their personal lives or their professions. Reflecting back on the individual cases, Interviewee F further discussed how Marissa Mayer or Theresa May faced the same scenarios. He showed a sense of sympathy as though he felt bad for the women rising during a time of crisis; when asked about his reaction, he stated that he felt it *"unfair to put a woman a gravel road while men are taking the smooth paved path towards the same destination, no wonder women end up a bit more tough when arriving at the destination [leadership position]."* Interviewee G added that he could not find an example of a *"feminine"* CEO.

The very compelling conversation was advanced when In-

terviewee H appended her point of view as woman in leadership. In a male-dominated workplace she reflected on her own path to management, with a cunning smirk on her face, she established that women are in charge of their own path to management. Stating that *"we [women] are conditioned to believe that men work against us and that we are less or not respected"* she explained that how one is perceived in the workplace is simply up to them. If one wants to be respected, he or she must demand that respect; *"I'm not sure if this is my Italian upbringing, but in Italy, how you walk down the street defines how you'll be treated."* She continued to expand on her experiences working with women, saying that her very direct mentality did not often fit to their working styles and that men often appreciated her straightforwardness, furthering the comment with *"when in Rome, do as the Romans do."* Interviewee H wanted to enter a male-dominated workplace and accepted the conditions corresponding with it; she said that if she were to be herself *"dancing, smiling and funny"* she would most likely lose her credibility in the workplace. Interviewee H is an exact depiction of the women who acquire *"masculine"* traits to achieve the managerial positions. While she may not want to realize or accept this fact, she exhibited the main four stereotypes associated with stereotypical men in successful managing positions of companies- *decisive, dominant, feelings not easily hurt and able to separate feelings from ideas* (M. K. Ryan et al., 2011).

This age faction was very adverse in their responses to the modern corporate world in regards to female leadership. Having experienced a lot more personally and naming numerous examples in their personal environments, all four exhibited the need for conversation of female challenges in ascending to management positions. Realizing that many of the women in their surroundings either took on a leadership position by changing their own methodologies in the workplace or by taking an opportunity presented to them, either through quotas or a glass cliff position.

4.3.3. The mature generation

Defined as the seniority of the participant assemblage, these Interviewees ranged between the ages of 46 to 60. Characteristically bringing more experience and insight to the conversation about gender discrimination in the workplace and the existence of the glass cliff. Each interviewee provided different insight on their career paths and the experiences they encountered along the way. With three women in management positions and one man, this faction had the most in depth and reflected answers also noting that the interviews with this faction took nearly twice as long as those in the intermediate phased group.

Opening the discussion with the differences in paths to leadership between men and women, all four interviewees agreed that there were explicit disparities. Interviewee K described *"the path for men is just very implied, it is nothing new and they know what is expected of them and how to deal with hurdles upon entering and rising in the workplace."* While she reflected that her comments were quite generalizing she contended further stating *"they [men] don't overthink all the problems they could face... it's a very direct culture, also in their path to leadership... if you do X, you get Y."* Interviewee M described the different approaches as transactional and transformational explaining that *"[men] use the transactional approach... wanting to achieve certain goals and each member of their team do their something that will result into reaching that goal... often not engaged with their team and their work."* Whereas, *"[women] use the transformational approach... the focus is more on how the goal is achieved, being very engaged with their team and their work."* Both women who have climbed the corporate ladder themselves, compared their paths to the journeys of the men around them. There was a serenity and composure in their responses, looking into the past with a sigh of relief that it's over, at least for now. Interviewee M added profoundly *"sometimes, you have to be the rock the river"* when asked about her tranquil answers. Interviewee J, the only man in the round, added that *"there is a choice between career and motherhood, women who become mothers appear to take considerably longer to reach leadership status in organizations."* Interviewee K reflected on the different phases of her career, starting off as a freelance architect in her mid-thirties, she confronted many prejudices about her work as a woman however, now that she has created a name for herself and a clientele, she rarely faces any gender discrimination upon a new project offer. All four of the interviewees are in management positions are have a hindsight view of their paths to their positions which extends the points of view described to be more reflective than futuristic and gives an insight into the paths transformed of how their climb was successful.

Addressing quotas in the management positions stirred Interviewee K to talk about equal opportunities in the workplace. She stated that *"quotas stand for opportunity and that's what the conversation needs to be about, women need to have the opportunity to get a position and since the foundation of society currently puts their management abilities behind those of men- they are needed as the stepping stone of progress."* While she hopes to agree that one-day society will be able to *"appoint on individual merit"* as Interviewee J pointed out who believes that quotas *"can lead to the appointment of a less suitable candidate,"* she doubts that certain male-dominated fields can give women the proper opportunity without certain regulations in place to do so. Working as a freelance architect, Interviewee L, did not encounter many quotas beside the admission rates during university however, she felt compelled to share that *"quotas are not the ideal way to promote women into workplaces however, in certain industries, they are needed for women to get their foot in the door."* She described how her freelance work does not call for quota positions and that a quota would most likely hurt her reputation instead; however, acknowledging that STEM positions, especially in upper management still *"majorly lack women in their leadership and that quotas maybe the only way to change the dinosaur foundations at hand."*

This prior discussion initiated the conversation into when women are hired into leadership positions, especially into glass cliff positions. Interviewee M opened up about a close

friend who was hired to *"clean up the mess"* when the company ran into financial trouble and as the numbers shifted from red to black, her senior position was no longer required. Despite her success in the company's mishap, the acknowledgement for her work was undeniable, more importantly, the company had several new hires in the following weeks all male in similar leading positions. Interviewee K reflected on her team experiences, where the more critical tasks were usually delegated to the woman in the group and how it was often the woman who was chosen to approach a delicate conversation if there was an issue at hand, this was if the confrontation went bad it would not fall back on the male members in group. She furthered on the public image in modern society, comparing politicians such as Theresa May in her post Brexit election and the Icelandic Prime Minister Johanna Siguroardottir who was elected after an extreme economic downfall. She then stated that Hollywood creates glass cliffs every day and expects the more of female actresses than male actors vulgarly proclaiming *"... nudity, women are much more openly portrayed nude in film [than men] and all that is ever seen [of men] is the buttocks."* She continued the conversation into the Me Too Movement (Appendix 1) explaining how *"the women who experienced any type of assault are put in class cliff positions where they have everything to lose to and not much to gain in stepping forward [about the assault]."*

Interviewee K finished the interview with a powerful message: *"... true confidence is the only solution to this problem, trusting in each other is essential to creating the greater change that so direly needs to happen in the working world. If women were to truly come together, I think we would be an undeniable force that could move mountains. We can play our strengths and work together... just truly embracing our femininity and working alongside men to make something even bigger than any gender issues that can be discussed in today's modern society."*

4.4. Post-interview

The 13 individuals who offered to partake in this research provided very different insights on their opinions and experiences in respect to women in leadership. Every interviewee added value to conversation and created the means for discussion, being open to being challenged and being asked about uncomfortable situations. There were unequivocal differences in the interviews done in person versus over the phone or written responses, even in the mere side discussion that arose from the questions. Nevertheless, identifying the existence of the glass cliff was critical and successful in these interviews and will be further interpreted in the next chapter.

5. Discussion

The following discussion aims to review the literature analyzed in the background chapter of this thesis. The background chapter of this thesis provided the fundamental principles that are needed to continue the conversation about female leadership and the opportunities facing women rising in the corporate world. After having predefined the different associations that coincide with the glass cliff phenomenon, the paper investigates the psychological phenomena that surround the creation of glass cliff situations in the first place. This directed the requisite to conduct personal interviews with a diverse range of people in different paths of life, reflecting on their own experiences and opinions about female leadership. Since the glass cliff is not caused through a single source, the substantiation of the surrounding existing research was furthered on in the preceding literature review and extenuatingly discussed in the interview sessions.

5.1. Major findings

This chapter distinguishes the major findings of the research conducted for this thesis. It will highlight the variety of gender discrimination facing women in the working world today. From the glass escalator to understanding the groundwork that creates glass cliff positions, this chapter will highlight how the main causations- implicit theories, TMTM, TCTF, status quo bias, stereotypes, sexism and group dynamics- contribute to the existence of the glass cliff.

5.1.1. Glass Escalator

Discussing the path to leadership with the interviewees alongside the precipitative literature review, the differences for men and women on ascending the corporate ladder are protrusive. From the enduring existence of the glass ceiling which continues to restrain women from corporate promotion due to male preconceptions and in reprehensible stereotyping of women (Hefferman, 2002) to the glass escalator imparting more leadership opportunity towards men working in female-dominated fields by receiving management positions in occupations such as education, nursing or social work (Broner et al., 2013, September 30). As a teacher in Cambridge, Interviewee E described her institution as female dominated and despite her role as head of her faculty she voiced that her *"senior leadership is very male dominated and there is explicit sexism."* She continued to explain that she felt the promotion process hired several premature individuals [men] before taking into account a woman for the position.

Women facing obstacles on the way to leadership positions was a common matter in the discussions with interviews. The mature group interviewed approached the explicit divergences in the path most eloquently, coming to the conclusion that *"[for men] the path is just very implied... they know what is expected of them... and they don't overthink all the problems they could face"* (Interviewee K). However, as Interviewee F pointed out, is important to acknowledge that men might not have to overthink the problems they could encounter because those problems simply do not exist for them. Essentially, that is what the glass ceiling, glass escalator and glass cliff illustrate; as women are entering new frontiers of managerial leadership, they face new, imperceptible barriers that restrict their capabilities.

5.1.2. Creating the glass cliff

The glass cliff phenomenon states that a woman is more likely to rise in leadership during poor company performance (M. K. Ryan & Haslam, 2005, M. Ryan & Haslam, 2007) and that crises cause society to look toward female leadership (M. K. Ryan et al., 2011). When introducing the glass cliff phenomena to the thirteen interviewees that partook in this thesis research, a plethora of insights and points of view were shared. These narrowed down to the impact that gender stereotypes have on societies point of view of female, overt sexism and the tainted stigma between female leaders.

Implicit theories

The implicit theories observed stated that there is an incompatibility in perception of beliefs about what it means to be a good leader and what it means to be female. (Agars, 2004; Eagly & Karaus, 2002; Heilman et al., 1989; Schein, 1973, Schein, 1975, Schein, 2001). More critically, the theories from Agars (2004) and Eagly and Karaus (2002) about gender and management are perilous as they are powerfully descriptive (Heilman, 2001) meaning that they create the foundation for the mannerisms in the workplace and the expectations of how women should act and perform. The thirteen interviews conducted all considered the misrepresentation of male characteristics in leadership and furthered on how society is impacted by the misperceptions. This included touching base on associations like the TMTM and TCTF correlations.

Think Manager—Think Male

After Schein (1973) proved that there is a direct correlation between the perceptions of what it means to be a good manager and what it means to be a woman, many researchers have replicated and furthered the research existing. Including Eagly et al. (1992) who assessed the mannerisms of assertiveness between men and women, the man was seen as a leader when being assertive and the woman was seen as pushy. As Interviewee C described the stereotype of women CEO's as *"sassy and smart... with court heels..."* which supports the existence of the TMTM principle since men are not judged negatively on the same actions a woman may do, including the point of workplace attire. As Interviewee K explained that *"when someone [hiring manager] views a feminine woman they do not associate manger styled qualities with her"* hence they will consider her differently when deciding the next steps in leadership.

Think Crisis—Think Female

Think crisis think female is a foundational element of the existence of the glass cliff (M. Ryan & Haslam, 2007). Confirming that women are more desired in a management position during a company's time of poor performance, these risky positions hold more potential for a woman to join the ranks of leadership (M. K. Ryan et al., 2011). The interviewees most frequently referred to the TCTF principle when relating it to women they had seen take on these positions like Marissa Mayer and Theresa May. Interviewee M and Interviewee I shared stories about friends that had experienced this phenomenon at work, where women were advanced in their careers because of a carping situation within the company. Interviewee F referred to these women as the *"daring angels"* who were willing to the job that others [men] were not.

Status Quo Bias

If something is not broken, there is no need to fix it; this statement is groundwork of the status quo bias which specifies people's reluctance to change something if there is no immediate need for it (Samuelson & Zeckhauser, 1988). The Status Quo Bias was immediately identified by Interviewee J who stated that the *"most dangerous sentence in the English language is "because we have always done it this way.""* He related the western corporate culture to management philosophy as whole, which made him understand the how status quo bias is applicable to issues women face when rising to leadership.

In a historical context, society has been heavily influenced by male-dominated leadership which has instinctively created a masculine stereotype about leadership (Dasgupta & Asgari, 2004; Eagly, 2007). This prejudice was previously confirmed by Schein (1973), Schein (1975) studies, where he analyzed how gender stereotypes affect the perceptions of what it means to be a good manager. His study revealed that the descriptive qualities associated with men were congruent with the descriptive terms associated managers; whereas, the terms associated with women were not associated with being a good manager. The impact that these stereotypes have on the perception of women in the workplace was extensively identified in the conversations conducted. The younger generation disclosed how they feel starting their career paths in terms of what is expected of them, fully aware of the likelihood that they will have to work harder than their male colleagues to receive the same amount of respect or leadership opportunity. Furthermore, Interviewee H represents the common case of women taking on *"masculine"* traits to succeed in a male-dominated environment, which can be a subconscious or conscious coping mechanism to react to the discrimination. This sentiment places women in a lose-lose situation, if they want to be good managers then they have to give up on their femininity vice versa insinuating that to be a *"good"* woman one cannot be a good manager.

Stereotypes

While discussing the stereotypes associated to leadership and women adapting to the masculine workforce, politics quickly made itself into the conversation. Interviewees noted that the stereotypes can impact how a woman executes her role, for instance Angela Merkel, the chancellor of Germany,

who embodies relatively masculine characteristics and is defined as a powerful leader. Alternatively, Jacinda Ardern, prime minister of New Zealand, encompasses a more feminine approach in leading her country and delegating discussions, this included having a child during her time in office being the first woman as head of a country to do so in thirty years[14]. Whether a woman succumbs to the stereotypical behavior of managerial roles can be dependent on the industry or background surrounding the position at hand. To offer another insight, one can attempt to understand why these women present themselves so differently by referring to Geert Hofstede's model of cultural dimensions. As seen in Figure 2 Germany has a very high score in the masculinity dimension tied with the United Kingdom and New Zealand falling eight points lower. When comparing the political leaders, Merkel and May have very different leadership characteristics to Arden; they come across to the public as very masculine women and *"are characterized by strength and resilience"* (Interviewee I). Hofstede describes masculine societies as *"People rather "live in order to work" and draw a lot of self- esteem from their tasks. Managers are expected to be decisive and assertive. Status is often shown, especially by cars, watches and technical devices."*[15] Hofstede's definition of the masculine society is an undeniable account of the stereotype associated to men in the modern western world. Interviewee J's comment that he found *"female leaders to be more ruthless than male leaders"* is therefore justifiable if that is how society illustrates masculinity.

Women rising into leadership face not only the prejudice within the working environment but an entire societies perception of what management looks like. After all, one could argue that in most governments there is a need for an internal reflection of the male to female ratios in place. According to the UN Women Permanent Missions calculation, as of January 2019, there are 11 female Heads of State and 10 women are serving as Heads of the Government.[16] Out of 195 countries, that is a bare equivalent and the women taking on these political leadership positions vary immensely in character and leadership style.

Hostile and benevolent sexism

Unfortunately, hostile sexism in the workplace has not been eradicated therefore, it is one of the many facets adding to the existence of the issues that women face in the workplace. There are *"dinosaurs"* in certain industries that will hinder the progression of women into leadership until the end of that era. Hostile sexism is a blatant causation of the glass cliff; it implies that some women appointments in the workplace have the intention to disadvantage a woman and to see women fail while overcoming the hurdles of the corporate world (M. Ryan & Haslam, 2007). Interviewee A's presumption that *"men believe they are superior without knowing their [women's] qualities first"* insinuates that as a 20-year-old medical student, she is still confronted with prejudiced views against women in the workplace. While the subtler art of discrimination is conveyed through benevolent sexism, where the woman is rewarded for her work but to a position that sets her up to fail (Cuddy et al., 2004; Glick & Fiske, 2001). While none of the interviewees reflected on having experienced benevolent sexism themselves, they did mention that in public positions such as political roles come with a substantial amount of sexism.

The times are changing, many industries are opening their eyes and arms to the assets women can be in companies today. While the majority of companies maintain a male-dominated upper management, there are more women in leadership positions than ever before (Catalyst, 2009; US Bureau of Labor Statistics, 2006). Additionally, there is institutional regulation pushing for advancement of women to boards like the California Senate Bill 826 (SB 826) necessitating a female board member on every board of directors in a public company. As Interviewee K had said *"quotas stand for opportunity... they are needed as the stepping stones of progress"* which agrees with Dasgupta and Asgari (2004) which states that the association of leadership and masculinity may be changed with higher representation of women in leadership positions. Having a quota in place for hiring standards can be misleading as well, whether it be in hiring regulations or benevolent methods of promotion. Interviewee K gave an example: if company A that has fifty percent women employed, where fifty percent of those women are secretaries and the board of directors consists of only men the *"opportunity for women to rise into leadership positions would still be minimal"*. Whereas, if company B has two women on a board of six members however only has thirty percent women in the organization, company B would be considered as "more egalitarian" than company A in terms of possibilities for women to rise to leadership positions. The argument between companies A and B makes it difficult to understand how beneficial quotas can be in the workplace, more importantly because they can negatively impact the social perspective of the women hired as they might not be deemed as hired by merit which could have negative reprimands. Nevertheless, Interviewee L said it well *"quotas are not the ideal way to promote women into workplaces however, in certain industries, they are needed for women to get their foot in the door."*

In-group favoritism

Interviewee J turned the conversation by adding the social factor that aligns to ones need to feel accepted and significant. *"Men develop cliques... around soccer, football and beer. If you are not part of this circle you are a threat and you do not receive the support of the group"* which is where a different sense of comradery is built that is therefore naturally reflected in the workplace. Just like women, the social factor matters immensely. If a woman feels threatened socially, she will go out with her friends and demote the person who is

[14] New Zealand PM Jacinda Ardern's baby premieres among world leaders.
[15] Country Comparison, n.d..
[16] Facts and figures: Leadership and political participation, n.d..

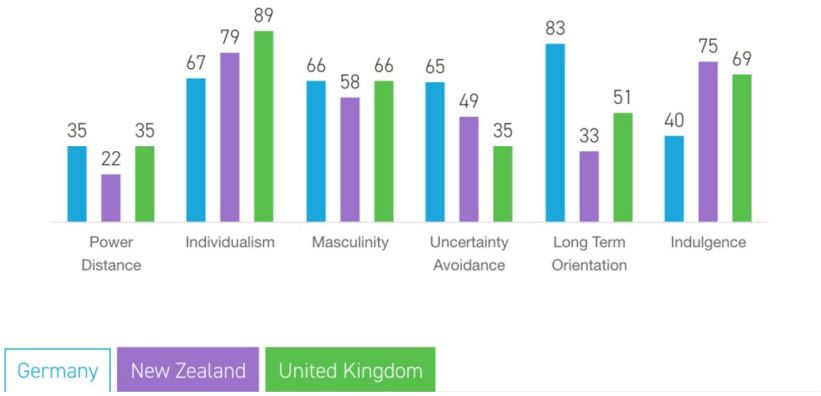


Figure 2: Hofstede model of cultural dimension comparing Germany, New Zealand and the United Kingdom (Hofstede, 1967)

making her feel vulnerable, the gossip then leads to exclusion and judgement from the group which if the person is in the vicinity can immediately be observed. As Interviewee I phrased it, *"by putting another person down in conversation, they are protecting their social position."* The same situation happens in the workplace, if a woman in an authoritarian position feels threatened she will lash out in a way that deems fit for the time and place. However, this does not explain the overall sentiment that women are often reluctant to help another woman progress in her career. Women are skeptical of other women's abilities as well as their own, whether it is a matter of insecurity or inferiority is undefined but Interviewee H argues that it is women perception of themselves that hinders their further performance. Interviewee K described it as *"lack in true confidence,"* and Interviewee M believes that some *"women want you to climb the ladder the same way they did... experience what they experience."* Moreover, Interviewee G explained it is *"unfair to put a woman a gravel road while men are taking the smooth paved path towards the same destination, no wonder women end up a bit more tough when arriving at the destination [leadership position]."* It is no longer a one sided discussion where men see women a certain way, it has become a universal perception that women are different to the qualities needed to be a good manager. Understanding why women can be so rancorous towards other women is another subject that can be further researched as it directly correlates to the impact that gender stereotypes can have on the progression of women in the workplace.

5.2. Meaning of findings

After having thoroughly analyzed the literature surrounding its existence in combination with the participation of the individuals involved in the qualitative research, the major finding of this thesis is the existence of the glass cliff and how it is created. By combining the insights collected from the interviews with the existing research, this discussion will be based on the glass cliff's existence and how it is addressed in the workplace, moreover, how it should be addressed in the future.

The glass cliff is not caused by one mere feature of modern societies perspective on women in the workplace, it is a combination of a plethora of different theories and biases that when collectively analyzed portray how glass cliff positions are created. It is imperative to realize that the *how* is directly correlated to the *when* of glass cliff positions creation, when certain themes are observed in the workplace such as implicit theories, biases regarding gender stereotypes and the status or unabashed sexism, the likelihood that these aggregations create leadership hurdles for women to advance is much higher. Per contra, identifying *why* glass cliff positions was indubitably the more challenging aspect of the research conducted, it summoned down to surrounding principles. The existence of the outside factors such as the biases, are the forefront of why the glass cliff exists, understanding why those factors still exist was not withstanding in the research conducted.

5.2.1. The path to leadership

Parsing into the causational factors of contention for women in modern society working towards the managerial roles revealed that the glass ceiling, glass escalator and glass cliff create a triad that encompasses the different struggles that women experience while climbing the corporate ladder into leadership positions as seen in Figure 3.

The triad identifies the core pillars that cause complications for women working their way up into leadership positions within an organization. To assume that she will encounter one of the three principles on the path to leadership would be anaberration, however if she does encounter an obstacle because of gender on the way, the likelihood that it is one of these three principles is prodigious.

The path to leadership for men and women was proven to be different before the research of this thesis further confirmed it. What researchers identified as the glass ceiling, was considered as a difference in approaches towards hir-

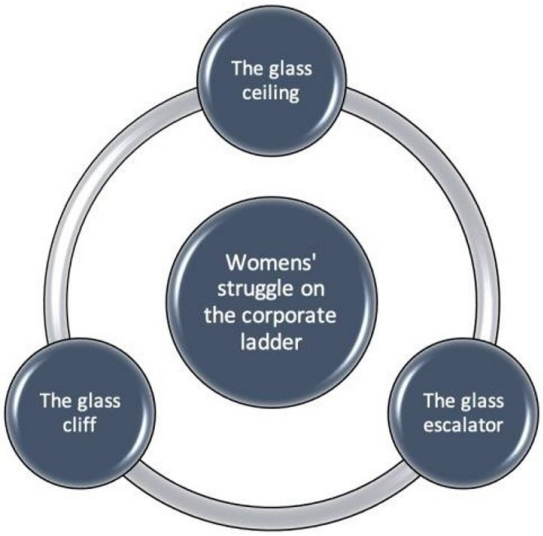


Figure 3: The hurdles women encounter on the path to leadership (Data compiled by author)

ing organizational management. As it was nicely put by one of the interviewees, the paths can be differentiated in the approaches taken- men with a transactional approach and women with a transformational one. Furthermore, the glass escalator was confirmed by one of the women working as a teacher, a female- dominated industry where the leadership is still very male dominated despite its faculty existing of a majority of women. The glass cliff as defined by Haslam and Ryan was confirmed through a profusion of examples and theories provided by the participants of the research, reflecting on personal experience and expertise as well as referencing larger scope positions seen in the media headlines of today.

5.2.2. The effect of Stereotypes

A further finding of this thesis paper and a critical attribution to the creation of glass cliff positions regards the stereotypes surrounding gender roles and managing capabilities. As first discovered by Schein (1973), Schein (1975) then further confirmed by a substantial amount of researchers including Haslam and Ryan, there is a direct correlation between the perception of genders and the ideology of what makes a good manager. This is the underlying foundation that creates the chasm between men and women when addressing progression into organizational leadership. Several interviewees agreed that the misperception of how women are and what they are capable of, is the driving point to extant hindrance of women to managerial positions. Historically speaking, the male- dominated culture of leadership shaped these perceptions of quality managing skills being associated to masculine character traits (Schein, 1973). This further includes women's perception of themselves and towards each other; women are also affected by the gender discrimination showing that they do not necessarily believe they can break through the triad principles themselves therefore falling victim to the discrimination. Women have the choice of adapting to the expectations of masculine leadership qualities or they can challenge their path by adopting a more feminine approach or *"transitional"* as Interviewee M called it. If women were to hold themselves in a different limelight, they may further change the stigma of the current workplace, which was the perspective that Interviewee H was providing. She claimed that it's a matter of personal choice on how the workplace treats one, this includes the matter of opportunity. Lastly, women can be malicious towards each other in the workplace, in attempts to conserve their own image or position. This was discussed several times, where some of the participants noted that gender discrimination is not a matter of men and women, it is fueled by both sides and needs to be changed from both sides as well.

5.2.3. The impact of quotas

Withal, the surfeit conversations about quotas have to be considered as major finding of this research. While the focus was not on the effects or perceptions of quotas for female representation in the workplace, the participants heavily reiterated their importance. It was proven that the higher representation of women in leadership positions, may lead to changes in the stereotypes currently surrounding leadership and masculinity (Dasgupta & Asgari, 2004). Further concluding from this research paper, that quotas are the necessary stepping stone in certain industries to advocate women getting into the senior ranks of management. This surrounds the extensive conversations about institutional regulations such as the SB 826 requiring female board members on public board of directors registered in the State of California, United States (SB 826). This was one of the more controversial aspects in the research conducted but it proved to be essential in understanding the reasons of difficulty for women climbing the corporate ladder and for the creation of glass cliffs.

The major findings in the research conducted confirmed the literature reviews' perspective on nearly entirely. It verifies the current situation of what the path to organizational leadership looks like for a woman in the 21st century and it provided the insights from three contrasting reference groups: the younger generation, the middle generation and the mature generation. By considering the different phases of life, it was evident to see how the past has shaped the current situation and how the current situation will change future. The topic of female leadership is very efficacious and its forefronts will see immense changes in the next decades, on an institutional level and social perspective.

5.3. Relation to prior work

The amalgamation of the research conducted for this thesis with the prior existing research provides a benefit to understanding the causation and existence of the glass cliff. By summarizing the works of M. K. Ryan and Haslam (2005), M. Ryan and Haslam (2007) in proving the existence of glass cliff positions alongside the foundational elements of Schein (1973), Schein (1975) studies that proved the stereotypical bias of what it means to be a manager in regards to gender. The research that the qualitative interviews of this paper proved is the aftermath that the glass cliff phenomenon has on men and women in the workplace, moreover on society as whole. The interviewees shared their own encounters with gender discrimination and glass cliff positions, which thereby validates the ramification that corresponds to how female leaders are perceived today. By understanding the individual experiences that the interviewees reflected on, patterns in the different age assimilations were visible which provided the means for understanding the impact that the glass cliff has on the different phases of a career and investigating how the stigma towards women in leadership was created. A decade after Haslam and Ryan's findings, the effects of the glass cliff are still valid and present in the modern workplace which therefore sets the relation between the research conducted for this thesis to the prior existing work.

5.4. Validity of work

To assess whether the qualitative research was valid, this section will assess the four criteria of validity- objectivity, reliability, internal validity and external validity- to ensure that the research conclusive. First to assess the objectivity of the research, one must speculate whether the study is replicable. In this case, the study is replicable because the same interview question set can be used in another setting; however, the results of the research could vary dramatically. Due to the qualitative insights that the participants shared, any new participant of different demographic or experience set, will change the projections of the research. It would be important to note the differences in gender, age, education, profession and demographic background to further understand certain conclusions drawn out during the research process. To assess the reliability of the research, one must cultivate on the steadiness and measurement accuracy of the data collected. Anew, the answers collected from the participants were under different settings as some interviews took place face-to-face and others were on the phone or via email. These differentiating factors in collecting data naturally skew the measurement perception of the project as one cannot read the body language or interpret the subtle gestures over the phone in comparison to in person; more importantly, the person analyzing the research is ordinarily not an objective person. In terms of internal validity, the variables tested for were constant which allowed for general pattern identification in the results. This concludes that I was able to measure what I wanted to measure. Lastly, the external validity accounts for the generalizability of the results, beyond this context. The interviews were not detail oriented nor demographically inclined, meaning that the conclusions drawn up in this report general in the term of results and can be applied to several broader perspectives.

6. Conclusion & Contribution

Women taking on leadership positions in the workplace has been an imperative matter internationally for decades. Breaking free of the limitations placed on them, women are rising to vanguard the path to leadership for every woman to follow. This thesis was dedicated to understanding the glass cliff, thereby bringing to life Eleanor Roosevelt's quote *"A woman is like a tea bag: you never know how strong she is until she gets into hot water"*[17]. It rings true in its foundation as it precludes women's tendency to thrive during a time crisis, which is essentially what the glass cliff phenomenon states. By researching the root cause of the glass cliff, this thesis defined how, why and when these positions are established furthering on how to approach the transformation essential to changing the workplace in modern society.

This chapter will aim to summarize the key findings determined from this thesis. It encapsulates the literature review, research findings and coherent discussion in regards to the analysis. It will further depict the potential limitations regarding the existing information and provide insight into future research that can be conducted pertaining to female leadership positions and overall management styles in respect to gender.

6.1. Conclusion

When investigating the glass cliff phenomenon, it was easy to presume the different aspects that permit its existence. However, after thoroughly analyzing existing research followed by personally investigating the causations behind glass cliff positions, it proved to be obscure to separate the *how, why* and *when* glass cliff positions are created. Therefore, the following four paragraphs may be repetitive in its conclusions. Nevertheless, the culmination of how the future workplace should be and how the glass cliff can be approached remains intrinsically captivating.

[17] (Ayres, 1996, p. 199)

6.1.1. The research

This thesis was written to explore the causation of glass cliff positions in three measures. First, by understanding *when* glass cliff positions emerge based on analyses of historically held positions of women in the corporate world. Then, by examining existing studies and executing qualitative interviews, this paper evaluated *how* and *why* these glass cliff positions are created in today's society. By exploring these elements, the research aimed to conceptualize the ideal working environment for men and women in modern society and address how gender discrimination can be addressed in the workplace.

6.1.2. When Glass Cliff occurs

Glass cliff positions occur when an environment fosters gender discrimination towards women in managerial positions. *When* stereotypes about what it takes to be a manager appear, more specifically, the male qualities needed to be a good manager prosper (Schein, 1973, Schein, 1975) the stigma towards women attempting to rise to a leadership position is created. This can be strengthened through blatant and benevolent sexism in the workplace, that lays the groundwork to disadvantage a woman climbing the corporate ladder. Conversely, when group dynamics and status quo bias present themselves converse to female leadership, creating the glass cliff is inevitable since this stigma restricts women from achieving a promotion during a time of normal company performance. *When* these different associations are intact, the capacity for a glass cliff position is created.

6.1.3. How Glass Cliff occurs

The glass cliff phenomenon occurs through the harboring of gender discrimination in the workplace. By perceiving the stereotype of men being better equipped leaders (Schein, 1973) as true, one fuels the foundation that gives life to existence of glass cliff positions. The stereotypes, biases, unabashed sexism and group dynamics featuring women at a disadvantage when it comes to managerial positions are the acumen that grow the negative sentiment of women in leadership. The aggregation of this sentiment in the workplace makes it difficult for women to be promoted under normal conditions which therefore makes them the more qualified individual to take on a precarious position. This evolution of playing towards women's characterization of being understanding, helpful and sophisticated (Schein, 1975) generates the path of creating a glass cliff position. The think crisis-think female (Schein, 2001) association is quintessential to the existence of the glass cliff and explains how the positions are created.

6.1.4. Why Glass Cliff occurs

The why behind the glass cliff phenomenon maybe the transparent attribute to understanding its existence. Historically, leadership positions have been filled by men and modern society has been subconsciously conditioned to associate leadership qualities with the stereotype of masculinity (Dasgupta & Asgari, 2004; Eagly, 2007). *"The single most important hurdle for women in management... is the persistent stereotype that associates management with being male"* (Antal & Izraeli, 1993). The explanation given by Antal and Izraeli is the same conclusion drawn from the interviewees of the research conducted for this thesis, it's the historical context that has driven society to perceive leadership in a masculine manner.

6.1.5. Idealistic Workplace

When discussing the future of women leadership and the ideal working environment, the participants of the study reflected an interesting pattern. They used the word *nurture* and *supportive* in describing the environment they would prefer to work in, this specifically reflected on the type of leadership they wish to see. Both of those terms were characterized as female descriptive terms (Schein, 1973) and appeared to match to similar qualities defined as desired of managers regardless of the performance of the company (M. K. Ryan et al., 2011). This proves that the working environments desired by employees and fellow leaders is one that supports and nurtures the others, something that would be described as a feminine approach. Essentially, the research conducted presumes that if men and women were to act more *"feminine"* in their leadership styles, their colleagues would appreciate it. As Dasgupta and Asgari (2004) had hypothesized, that a higher representation of women in positions of organizational leadership could lead to changes in the associations between masculinity and leadership.

Essentially, by women stepping into the essence of their leadership skills and men granting the space for women to thrive in their fashion, the stigma currently held to female leadership will change. This requires both men and women to come together to stand for equal opportunity, embracing the qualities that make them different and more importantly, the same. It's easy to see differences but in retrospect, men and women are not as different as one might presume, more importantly they are both entirely human. It will take every single person to stand up and provide the means for equal promotion to one day eliminate glass cliff positions. Instead of making leadership opportunities a lose-lose situation for one party, by playing to each other's strengths, a win-win situation can be created in the workplace.

6.2. Limitations

To ensure a perceptive research analysis, one must consider the limitations of their research. The primal reservation this paper encountered would be the pool of participants interviewed during the qualitative study. Due to the restricted access to a certain demographic of people, the concoction of participants was based on the writer's access and the participant's willingness to contribute to the research. This resulted in the majority of the participants being European, more specifically, British. Furthermore, their backgrounds were relatively dispersed; to get a more vivid image of an

industry one may consider focusing on only that industry to gain more expertise to draw more precise conclusions. In regard to the gender balance, there were four men and nine women interviewed which reflects a natural shift in opinion, especially in regard to a topic where gender is the key factor of influence. The next study may consider only researching one gender or keeping the intake of research equal between the genders. Furthermore, the subjectivity of the interviewees can be undermined due to their demographic background and their relationship to the author. Due to a majority of the participants coming from a personal development background they may have a prejudice towards the topic of female leadership.

More objectively speaking, the sample size was small with only thirteen participants. A larger sample size will allow for more distinct, statistical conclusions to be drawn; however, a more quantitative approach may be considered for a larger sample size. Ultimately, the time frame of this research was also a limitation. In comparison to a yearlong study, the eight-week research period scurried the process of collecting information which therefore also restricted the access to certain participants.

6.3. Future Research

Future research regarding the glass cliff phenomenon has a vast amount of opportunity. Analyzing in depth what the driving factors are behind women accepting glass cliff positions, if daring to say, even enjoy them. Then discovering and analyzing the glass cliff positions in depth by including sample groups that have experienced the glass position, not just seen it from a third perspective. This would also provide further insight on the possible ways eliminating the glass cliff in the workplace with the commitment of men and women.

When generalizing into female leadership as whole, there is extensive research that can be conducted on the workplace in regards to how women treat each other in the workplace. Arguing whether glass cliffs, glass escalators and glass ceilings are the only issue hindering women's advancements into leadership, whereas the reluctance of women to help other women is not uncommon. Out of the eight women interviewed for this thesis, all eight stated that they had experienced negative female leadership in the past. Why would women not want to help other women? Is the fear of being out beaten or surpassed by a fellow woman bigger than the comradery that could propel them further? Further embarking on the dissonance many women associate with female networking events and the general perception of women amongst women is undiscovered. The idea behind women standing the in the way of each other's success is similar to that of a glass window; one woman has reached the top and could help another woman in their position, however chooses to look through the window instead of exchanging the knowledge both are exposed to.

Women's leadership will continue to be topic of discussion and research for the next decades. Eventually, there will be the first board of directors consisting only of women and there will be bountiful research analyzing their success or downfall. Furthermore, the social roles of women are changing in society and as the millennials arrive in the management positions, a new era of leadership will replace the current pillars of the corporate ladder.

References

Agars, M. D. (2004). Reconsidering the impact of gender stereotypes on the advancement of women in organizations. *Psychology of Women Quarterly*, *28*, 103–111.

Antal, A. B., & Izraeli, D. N. (1993). *A global comparison of women in management: Women managers in their homelands and as expatriates. in e. a. fagenson (ed.), women and work: A research and policy series, vol. 4.women in management: Trends, issues, and challenges in managerial diversity*. Thousand Oaks, CA: Sage Publications.

Ayres, A. (1996). *The wit and wisdom of Eleanor Roosevelt*. New York, NY: Plume.

Broner, C. K., Asselmeier, A., & Broner, C. K. (2013, September 30). *Men, Women & the Glass Escalator.* Retrieved from https://www.womenonbusiness.com/men-women-the-glass-escalator/

Bruckmüller, S., & Branscombe, N. R. (2010). The glass cliff: When and why women are selected as leaders in crisis contexts. *British Journal of Social Psychology*, *49*(3), 433–451.

Bullard, A. M., & Wright, D. S. (1993). Circumventing the glass ceiling: Women executives in American state governments. *Public Administration Review*, *53*, 189–202.

Bureau of Labor Statistics. (2005). *Women in the labor force: A databook.Washington, DC: U.S. Department of Labor.*

Catalyst. (2009). *2008 Catalyst census of women corporate officers and top earners of the FP500.* Retrieved from http://www.catalyst.org/publication/295/2008-catalyst-census-of-womencorporate-officers-and-top-earners-of-the-fp500

Center for Creative Leadership. (n.d.). *What Experienced Leaders Need to Know to Succeed.* Retrieved from https://www.ccl.org/articles/leading-effectively-articles/senior-level-trade-offs-what-experienced-leaders-need-to-know/

Country Comparison. (n.d.). *Retrieved from: https://www.hofstede-insights.com/country-comparison/germany,new-zealand,the-uk/.* Retrieved from https://www.hofstede-insights.com/country-comparison/germany,new-zealand,the-uk/

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2004). When professionals become mothers, warmth doesn't cut the ice. *Journal of Social Issues*, *60*, 701–718.

Dasgupta, N., & Asgari, S. (2004). Seeing is believing: Exposure to counter stereotypic women leaders and its effect on the malleability of automatic gender stereotyping. *Journal of Experimental Social Psychology*, *40*, 642–658.

Dreher, G. F. (2003). Breaking the glass ceiling: The effects of sex ratios and work-life programs on female leadership at the top. *Human Relations*, *56*, 541–562.

Eagly, A. H. (2007). Female leadership advantage and disadvantage: Resolving the contradictions. *Psychology of Women Quarterly*, *31*, 1–12.

Eagly, A. H., & Karaus, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review*, *109*, 573–598.

Eagly, A. H., Makhijani, M. G., & Klonsky, B. G. (1992). Gender and the evaluation of leaders: A meta-analysis. *Psychological Bulletin*, *111*, 3–22.

Equal Opportunities Commission (EOC). (2002). *Women and men in Britain: Management. Manchester, UK: EOC.*

Facts and figures: Leadership and political participation. (n.d.). *Retrieved from: http://www.unwomen.org/en/what-we-do/leadership-and-political- participation/facts-and-figures.*

Ferris, S. P., Jagannathan, M., & Pritchard, A. C. (2003). Too busy to mind the business? Monitoring by directors with multiple board appointments. *Journal of Finance*, *58*, 1087–1111.

Forbes, AllBusiness. (2018, October 4). *California Mandates Female Representation On Public Company Boards.* Retrieved from https://www.forbes.com/sites/allbusiness/2018/10/01/california-mandates-female-representation-public-company-boards/#7ade83ef1775

Fortune (n.p.). (n.d.). *6 Oldest CEOs in America in 2015.* Retrieved from http://fortune.com/2015/12/13/oldest-ceos-fortune-500/

Glass Ceiling. (n.d.). *Merriam-Webster's collegiate dictionary.* Retrieved from https://www.merriam-webster.com/dictionary/glassceiling/

Glick, P., & Fiske, S. T. (2001). An ambivalent alliance: Hostile and benevolent sexism as complementary justifications for gender inequality. *American Psychologist*, *56*, 109–118.

Haslam, S. A., & Ryan, M. K. (2008). The road to the glass cliff: Differences in the perceived suitability of men and women for leadership positions in succeeding and failing organizations. *Leadership Quarterly*, *19*, 530–546.

Hefferman, M. (2002). Exhibit A: The Female CEO. *Fast Company*, *61*(August), 58–66.

Heilman, M. E. (2001). Description and prescription: How gender stereotypes prevent women's ascent up the organizational ladder. *Journal of Social Issues*, *57*, 657–674.

Heilman, M. E., Block, C. J., Martell, R. F., & Simon, M. C. (1989). Has anything changed? Current characterizations of men, women, and managers. *Journal of Applied Psychology*, *74*, 935–942.

Higgins, M. O., & Kram, K. E. (2001). Reconceptualizing men to ring at work: A developmental network perspective. *Academy of Management Review*, *26*, 264–288.

Ibarra, H. (1993). Personal networks of women and minorities in management: A conceptual framework. *Academy of Management Review*, *18*, 56–87.

Judge, E. (2003). Women on board: Help or hindrance? . *Times*, *November*(11), 21.

Kaplan, S., & Minton, B. (1994). Appointments of outsiders to Japanese boards: Determinants and implications for managers. *Journal of Financial Economics*, *36*, 225–257.

King, E. D. (2006). Report from SIOP: Current industrial/organizational psychology perspectives of women in organizations. . *Equal Opportunities International*, *25*, 215–218.

Lead. (n.d.). *Merriam-Webster's collegiate dictionary.* Retrieved from https://www.merriam-webster.com/dictionary/lead

Leadership. (n.d.). *Merriam-Webster's collegiate dictionary.* Retrieved from https://www.merriam-webster.com/dictionary/leadership

Ragins, B. R., Cotton, J. L., & Miller, J. S. (2000). Marginal mentoring: The effects of type of mentor, quality of relationship, and program design on work and career attitudes. *Academy of Management Journal*, *43*, 1177–1194.

Ramsay, H., & Parker, M. (1992). *Gender, bureaucracy and organizational culture. in a. witz & m. savage (eds.), gender and bureaucracy*. Oxford: Blackwell.

Ryan, M., & Haslam, S. (2007). The Glass Cliff: Exploring the Dynamics Surrounding the Appointment of Women to Precarious Leadership Positions. *The Academy of Management Review*, *32*(2), 549–572.

Ryan, M. K., & Haslam, S. A. (2005). The glass cliff: Evidence that women are over-represented in precarious leadership positions. *British Journal of Management*, *16*, 81–90.

Ryan, M. K., Haslam, S. A., Hersby, M. D., & Bongiorno, R. (2011). Think crisis–think female: The glass cliff and contextual variation in the think manager–think male stereotype. . *Journal of Applied Psychology*, *96*(3), 470–484.

Samuelson, W., & Zeckhauser, R. J. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty*, *1*, 7–59.

Schein, V. E. (1973). The relationship between sex role stereo types and requisite management characteristics. *Journal of Applied Psychology*, *57*, 95–105.

Schein, V. E. (1975). The relationship between sex role stereo types and requisite management characteristics among female managers. *Journal of Applied Psychology*, *60*, 340–344.

Schein, V. E. (2001). A global look at psychological barriers to women's progress in management. *Journal of Social Issues*, *57*, 675–688.

Schmitt, M. T., Ellemers, N., & Branscombe, N. R. (2003). *Perceiving and responding to gender discrimination in organizations. in s. a. haslam, d. van knippenberg, m. j. platow, & n. ellemers (eds.), social identity at work: Developing theory for organizational practice*. Philadelphia, PA: Psychology Press.

Sexism. (n.d.). *Merriam-Webster's collegiate dictionary.* Retrieved from https://www.merriam-webster.com/dictionary/sexism/

Stereotype. (n.d.). *Merriam-Webster's collegiate dictionary.* Retrieved from https://www.merriam-webster.com/dictionary/stereotype#synonyms

Tharenou, P., Latimer, S., & Conroy, D. (1994). How do you make it to the top? An examination of influences on women and men managerial advancement. *Academy of Management Journal*, *37*, 899–931.

US Bureau of Labor Statistics. (2006). *Women in the labor force: A databook. Report 996.* Retrieved from http://www.bls.gov/news.release/pdf/atus.pdf

Williams, C. (1992). The Glass Escalator: Hidden Advantages for Men in the "Female" Professions. *Social Problems*, *39*(3), 253–267.

Junior Management Science 6(1) (2021) 60-80

Junior Management Science

journal homepage: www.jums.academy

Employment Protection Legislation, Youth Unemployment and the Role of the Educational System

Kimberly Klebolte

University of Oxford

Abstract

Research on the effect of employment protection legislation (EPL) on unemployment is extensive. However, results are ambiguous and were not able to show a clear pattern of how EPL is affecting labour market outcomes. Recent research has focussed on the effect of EPL on youth unemployment, linking higher protection to higher unemployment among young labour market entrants compared to their adult peers. Moreover, it is argued that EPL might not have a universal effect on youth unemployment but must be considered in an interplay of institutional factors, including vocational specificity. Based on these findings, this thesis provides a comprehensive assessment of the effects of EPL and vocational specificity on the labour market chances of young people compared to adults. As young people in particular often find their way into the labour market via temporary contracts, it is distinguished between EPL for regular and temporary contracts. A total of 28 OECD countries are examined from 1985 to 2013 using OECD data. In line with previous research, there appear to be no main effects of regular EPL or the vocational specificity on its own on the level of youth unemployment or the youth-to-adult unemployment ratio, but there is a positive effect of temporary EPL on the youth-to-adult ratio. This suggests that especially for young people deregulation of these contracts – contrary to the usual theoretical assumptions – can have a positive effect on their labour market situation.

Keywords: Labour market; EPL; unemployment; youth; education.

1. Introduction

Unemployment is a much-discussed issue in national governments and international organisations. Comparative Worldbank statistics show that most countries have seen an increase in youth unemployment in recent years and that youth unemployment is generally much higher than adult unemployment (Zimmermann et al., 2013) while there are clear differences in the unemployment rates of young people between 15 and 24 years across the OECD[1]. Especially labour market institutions and employment protection legislation are said to have a strong influence on unemployment. Employ-ment protection legislation (EPL hereafter) has very often been associated with high unemployment, but there is no consensus to confirm this link empirically and results seem to be inconclusive (Nickell and Layard, 1999). Nevertheless, EPL has not been lost sight of, but it is increasingly argued that strict EPL has a particularly negative impact on young people, as they are less likely to be hired, as employers have no references and information on their productivity and value to the company and thus fear high dismissal costs in the event of a poor fit (Esping-Andersen, 2000; Cahuc et al., 2004; Breen, 2005). Noelke (2015) and others show through the replication of prior research that strict EPL, however, is not to blame for high youth unemployment – at least not as the only factor. On the basis of these research results, the interconnectivity of different institutional factors becomes relevant and scholars argue that EPL does not have a clear effect on its own but depends on other contextual factors (Brzinsky-Fay, 2017). An obvious link here is the education system, which aims to prepare young people for the labour market and comes out strongly in past research as the institution that could make a difference. In addition to many factors such as investment in education, compulsory schooling or tracking, vocational specificity stands out as one crucial feature, since the link between schools and the labour market is most dominant. Signalling theory assumes that young people have a smoother school-to-work transition in countries with high vocational specificity and

[1] Blanchflower and Freeman (2000) provide a comprehensive review of cross-national differences and trends.

DOI: https://doi.org/10.5282/jums/v6i1pp60-80

a strong link between the education system and employers (O'higgins, 2001; Cahuc et al., 2013; Eichhorst et al., 2013). Here employers receive early signals about the productivity and capacity of potential workers so the fit to the company can be assessed better and more realistically. Since pupils in a vocational system are better prepared for the real problems and challenges faced in the labour market and get in touch with companies at an early stage, youth unemployment is expected to decrease (Quintini and Manfredi, 2009).

The most immediate departure points of this thesis are studies from Breen (2005) and Brzinsky-Fay (2017), which are the only ones working on the interlinkage of education and EPL. Breen assumes that strict EPL leads to high youth-to-adult unemployment, but that high vocational specificity can offset these negative consequences. Accordingly, the youth-to-adult unemployment rate should be highest for strict EPL and low educational signalling and lowest for weak EPL and high educational signalling - although Breen finds no existing countries combining the latter. Brzinsky-Fay suspects that EPL on its own has no effect on the youth-to-adult unemployment ratio and only becomes important in combination with other institutional factors. It is expected that strict EPL only has a negative effect on the youth labour market situation if vocational specificity is low. This thesis contributes to the literature and takes further the ideas of Breen and Brzinsky-Fay by improving on their work as not only EPL on regular contracts is examined but also the effect of EPL on temporary contracts. Moreover, due to new data availability, this thesis expands on both, but especially on Breen's study with regard to the time period investigated, covering a significantly longer period from 1985 - 2013.

The main expectation is in line with both that youth labour market situations can only be understood in an complex interplay of institutions: Regarding the main effects of EPL, I expect that regular EPL will have no main influence on its own (Noelke, 2015), but that strict regulations on temporary contracts will have a positive effect on young people, as employers will not be able to take uncontrolled advantage of these insecure working conditions to get cheaper and less protected workers. It is expected that the best labour market setting for young people in terms of unemployment reduction is one that combines low regulations on regular contracts with strict regulations on temporary contracts. In addition, the labour market situation of young people in countries with high vocational specificity, which creates a direct link between school and the labour market, is strengthened. Generally, an interaction of the individual factors is always assumed, so that neither EPL regular nor vocational specificity alone are expected to have a significant influence. Although the focus of this work is on the level of youth unemployment, adult unemployment is also included in the analysis using the youth-to-adult unemployment ratio as the second dependent variable to examine the different effects of EPL and vocational specificity on these two groups. Especially with regard to the distinction between regular and temporary contracts, it seems promising to include the ratio to capture different effects on adults versus youth, since permanent contracts are still the status quo for adults, while temporary contracts are increasingly becoming the norm for young people. Furthermore, the ratio can be used to proxy the overall state of the labour market. New results could be used to better understand the interplay between the regulation of regular and temporary EPL and vocational specificity with regard to future policy decisions in the field of labour market regulation or the design of education systems to reduce unemployment. Methodically, different OLS regressions are calculated with country-clustered robust standard errors in order to extract the main effects of both types of EPL and vocational specificity on youth unemployment and the youth-to-adult unemployment ratio as well as possible interactions between these key explanatory variables. Data for the two dependent variables as well as indicators for strictness of employment protection and secondary education vocational enrolment as a proxy for educational signalling are taken from different OECD databases.

In the following a literature review and the theoretical framework with hypotheses are given in part two. Part three covers the analytical strategy and contains information on the comparative research design, variables and their operationalisations and the statistical methodology. In part four, the descriptive results are reported first, followed by the results of the regression analyses and the robustness checks. Finally, a critical evaluation of this thesis with its limitations follows in the conclusion and an outlook on future research opportunities is provided.

2. Literature Review and Theoretical Framework

2.1. Employment Protection Legislation and Youth Unemployment

While many things can play a role in the institutional framework when it comes to explanations of unemployment, such as minimum wages, education and training or active labour market policies (Salvador and Leiner-Killinger, 2008), employment protection in particular is often associated with unemployment and job insecurity among workers – especially young labour market entrants (Jimeno and Rodriguez-Palenzuela, 2002; Esping-Andersen, 2000; Organisation for Economic Cooperation and Development, 2004; Breen, 2005). Employment protection legislation includes all the provisions on job security, e.g. severance pay or notice, which restrict employers' abilities to dismiss workers. The main function is to stabilise workers and their employment relationships and secure their jobs (Organisation for Economic Cooperation and Development, 2004). High EPLs thus generate costs for employers if they dismiss workers with permanent contracts (for an overview, see Esping-Andersen, 2000; Organisation for Economic Cooperation and Development, 2013; Gebel and Giesecke, 2011), so that fewer employment changes take place (Bukodi and Robert, 2007; DiPrete, 2002).

Even though EPL has often been associated with unemployment (Blanchard and Wolfers, 2000; Botero et al., 2004;

Di Tella and MacCulloch, 2005; Feldmann, 2009; Fialová and Schneider, 2009; Nickell, 1997), past research has not yet been able to find a common consensus regarding the impact of strict EPL on the aggregated unemployment rate (Nickell and Layard, 1999). Some studies find no influence at all or even a positive influence of EPL on the employment rate (Allard and Lindert, 2006; Amable et al., 2007; Belot and van Ours, 2004; Cazes and Nesporova, 2007; Griffith et al., 2007; Rovelli and Bruno, 2008). Some other studies were able to show a strong effect of strict EPL on turnover but could not find any effect on aggregated unemployment or unemployment rates (Kugler and Pica, 2008; Marinescu, 2009). A potential reason why past research might not have been able to find clear results is that there seem to be two mechanisms of EPL that work in opposite directions and thus possibly balance each other out (Bentolila and Bertola, 1990).

On the one hand, there is a negative effect of strict EPL for all those who want to (re-)enter the labour market. Employers will be less likely to hire new employees under strong regulations as they fear high dismissal costs if the employee does not prove to be a good fit. They will have to take into account future dismissal costs when hiring new staff, so that the benefits of hiring new staff are reduced and recruitment rates therefore decrease (Blossfeld et al., 2008). It is assumed that strict EPL restricts market flexibility and that employers do not hire new employees even in good economic times. On the other hand, however, strict EPL also has a positive effect for all those insiders who are already employed. With strict EPL redundancies are made costly with an aim to reduce the turnover, stabilise new employment relationships and protect workers from quick dismissals.

Studies show that both mechanisms are happening simultaneously as strict EPL reduces hiring and limits employee turnover on one hand but also prolongs existing employment relationships by reducing firing rates (Organisation for Economic Cooperation and Development, 1999, 2004; Kugler and Pica, 2008; Marinescu, 2009). Thus, a general theoretical and empirical relationship between EPL and aggregated unemployment could not be found (Nickell and Layard, 1999; Baccaro and Rei, 2007; Kahn, 2012), as it is a possible interpretation that the opposing effects of EPL outlined above have an equal impact on the aggregated unemployment rate (Noelke, 2015). As both mechanisms may balance each other out, it is the next step to look at different groups and not just at the aggregated unemployment rate to find out if the mechanisms might work differently for different groups. The decisive factor that emerged in the literature is the comparison between young people and adults. Strict EPL is cited in many studies as the main cause of integration problems of young people into the labour market and resulting youth unemployment (Esping-Andersen, 2000; Cahuc et al., 2004; Organisation for Economic Cooperation and Development, 2006; Breen, 2005; Boeri and van Ours, 2008).

While the protection argument for labour market insiders applies to both groups, the negative effect of decreased hiring rates might be particularly relevant for young people, so that they are more affected by these negative consequences than adults. In their school-to-work transition young people try to find their first job and take the first step into the labour market. However, if employers have high dismissal costs, they are more inhibited to hire new employees, which has a particularly negative impact on young people who always start as labour market outsiders when leaving education. Looking at the situation of young people, it can be seen that a reduction in job offers through strict EPL makes job search more difficult as the school-to-work transition is prolonged until the first employment is entered (Wolbers, 2007) and job opportunities for young unemployed people are generally worse in countries with strict EPL (Russell and O'Connell, 2001). Therefore, lower hiring rates might be more consequential for young people in their school-to-work transition than negative effects of higher dismissal rates for labour market insiders when there is less regulation. Based on this assumed relationship, there was consensus in international organisations (Organisation for Economic Cooperation and Development, 2004, 2006; European Commission, 2006, 2011; World Bank, 2007) and across disciplines that strict EPL increases youth unemployment in absolute terms and in relation to adults in OECD countries (Addison and Teixeira, 2003; Autor et al., 2006; Bertola et al., 2007; Blau and Kahn, 1999; Boeri and van Ours, 2008; Botero et al., 2004; Esping-Andersen, 2000; Kahn, 2012).

Although a large number of authors support the positive correlation between strict EPL and high youth unemployment, doubts are increasingly emerging that question the proposed link. Scholars from various disciplines and international organisations are increasingly critical of the assumption that EPL is a key driver of high youth unemployment. Noelke (2015) shows in his literature review and the replication of prior research in a comprehensive sample of Western European countries that a number of studies linking EPL with high youth unemployment are statistically not robust. He shows that the prevailing consensus that strict EPL is to blame for high youth unemployment in the OECD is not backed up with comprehensive evidence and agrees with previous findings that could also find no consistent effects (Dieckhoff and Steiber, 2012; Kahn, 2010; Organisation for Economic Cooperation and Development, 1999). Brzinsky-Fay (2017) shows with QCA that the claim that high employment protection legislation itself leads to high youth unemployment must be refuted (Brzinsky-Fay, 2017), as must Gebel and Giesecke (2016), who argue that more stringent employment protection legislation for regular contracts increases youth job insecurities, but does not lead to an increase in youth unemployment (Gebel and Giesecke, 2016) leading to the following first hypothesis:

> H1: EPL on regular contracts has no significant impact on the level of youth unem-ployment and the youth-to-adult unemployment ratio

While focussing on young people and their employability in the context of EPL, the different forms of EPL - regulations on regular and on temporary contracts - must also be considered more closely, as this distinction has been overlooked in

many studies so far (Gebel and Giesecke, 2016; see Dieckhoff and Steiber, 2012 for an exception). A moderately low correlation with 0.31 in an analysis of Gebel and Giesecke (2016) however, shows that both indices measure different features of EPL and that therefore a distinction should definitely be made. Permanent or regular contracts have no end of employment while temporary contracts have a predetermined date of expiration (Gebel and Giesecke, 2011; Organisation for Economic Cooperation and Development, 2004). For employers to not simply replace permanent contracts with temporary contracts, the use of temporary contracts has traditionally been very restricted with limits on the cumulative duration of these contracts or the number of temporary contracts that can be concluded in succession (Noelke, 2015). Generally, an employment with a temporary contract is more insecure and riskier because the end of an employment relationship is already foreseeable, and one has to take care of a new job during employment to not slip into unemployment after the end of a temporary employment.

In theory, there are two mechanisms of temporary contracts (Giesecke and Groß, 2003): The 'integration view' emphasises the positive advantages of temporary contracts, as they can stimulate the flows on the labour market, increase the number of hires and provide a first step into the labour market. The 'segmentation view', however, stresses a higher risk of job loss due to shorter employment contracts. The criticism of deregulation of temporary contracts seems to predominate the literature, which convicts a deregulation of temporary contracts while maintaining the same strict EPL for regular contracts as a main reason for high youth unemployment (S. and Dolado, 1994; Blanchard and Landier, 2002; Cahuc and Postel-Vinay, 2002; European Commission, 2011). This already shows that the regulations of regular and temporary contracts should always be considered in combination. While job security provisions in form of EPL on regular contracts have remained relatively strict over the past decades, temporary contracts have been gradually deregulated in many OECD countries (Organisation for Economic Cooperation and Development, 2004; Gebel and Giesecke, 2011). Schönmann and Clauwaert (2012) show a trend during the recession that more and more countries have liberalised and deregulated their labour markets for temporary contracts and new forms of temporary contracts have been created that are often less protected and explicitly targeted at young people (Schönmann and Clauwaert, 2012). The 'partial deregulation' (Blanchard and Landier, 2002) leads to the fact that with high dismissal costs for permanent contracts but low regulations for temporary contracts, there are no longer any economic and monetary incentives to convert temporary contracts into permanent ones. Young people are thus in a 'trap' of frequently changing temporary jobs, which are usually paid less and only limited further training is possible (Quintini and Manfredi, 2009). It is therefore argued that social inequalities are increased by the spread of temporary contracts (Gash and McGinnity, 2007) without creating and improving new labour market opportunities for young people, resulting in individual career risks and impairing social mobility (Barbieri, 2009; Cazes and Tonin, 2010; DiPrete et al., 2006; Gebel, 2009; Giesecke and Groß, 2003; McGinnity et al., 2005). In addition, employers are given incentives to replace permanent contracts with temporary contracts if partial deregulation happens (Blanchard and Landier, 2002; Cahuc and Postel-Vinay, 2002; DiPrete et al., 2006; Kahn, 2010). Moreover, an increasing difference between dismissal costs from regular to temporary contracts leads to a higher flow of temporary jobs into unemployment (Blanchard and Landier, 2002, see Centeno and Novo, 2012 for evidence).

It should be pointed out that young people in particular are disproportionately affected by this type of contract and thus also by the consequences of constant deregulation. Therefore, it is interesting to not only look at youth unemployment but also at the youth-to-adult unemployment ratio, which provides insights into how far young people are more or differently affected by changes than adults. Although time-unlimited contracts are the status quo of employment relationships among adults in the labour market (Noelke, 2015), it turns out that the increasing number of temporary contracts for all entry levels jobs affects disproportionately young people, for whom this type of temporary contract has become the norm in some countries (Organisation for Economic Cooperation and Development, 2014). Young people in particular are thus automatically exposed to a higher risk of unemployment due to the nature of their employment contract, as the positive insider effect of strict regular EPL does not apply to them. Furthermore, the destabilising effect of temporary jobs for employment careers and the failure of deregulation to lead to job growth for young people is stressed (Barbieri, 2009; Giesecke and Groß, 2003; Polavieja, 2003; Scherer, 2005; Barbieri, 2009). Gebel and Giesecke (2016) show across countries and time that overall partial deregulation was not successful in reducing the risk of youth unemployment, but that instead the youth temporary employment risk was increased (Gebel and Giesecke, 2016).

> H2: Strict EPL on temporary contracts is decreasing the level of youth unemployment and the youth-to-adult unemployment ratio

It is expected that the effect of deregulation of temporary contracts depends on the level of regulation of regular contracts, so the link of both types of EPL is crucial. Noelke (2015) confirms this expectation and shows empirically that youth unemployment rates rise when a country combines high regulations for regular contracts with a deregulation of temporary contracts, which is consistent with theoretical macroeconomic models (Blanchard and Landier, 2002; Cahuc and Postel-Vinay, 2002). What counts with regard to general unemployment but especially youth unemployment is therefore the link between regulations of regular and temporary contracts. Thus, both types of EPL will be examined in this thesis. The assumed best-case scenario for young people is having low regulations on regular contracts, as employers do not have to fear high dismissal costs if they want to employ labour market entrants on a secure regular contract and simultaneously high regulations on temporary contracts, as

a protection against being employed on insecure short-term contracts over and over again without planning certainty and with lower salaries. For adults, high regulations for temporary contracts are also important for this reason. For them, however, strict EPLs on regular contracts are better, as they are mostly labour market insiders and thus protected from dismissal by many rules and high dismissal costs.

> H3: The level of youth unemployment and the youth-to-adult unemployment ratio is lowest in countries with low EPL regular and high EPL temporary

2.2. Linking Employment Protection Legislation and Education

The theoretical and empirical backgrounds on EPL show that the relationship between EPL and youth labour market chances is not always clear, so that it seems logical to consider other factors that might influence this interaction. The interplay of different factors seems to have been somewhat neglected in the literature so far (with the exceptions of Breen, 2005 and Brzinsky-Fay, 2017). From a very detailed comparison of several countries, Zimmermann et al. (2013) conclude that a variety of factors influence the labour market situation of young people in a country, including the interplay of economic growth, labour market regulation or education and training systems. Avdagic (2015) also argues that the sole liberalisation of employment protection laws by governments will remain unsuccessful in the fight against unemployment. In line with these interconnectivity arguments it is argued that EPL might not have a universal effect on youth unemployment but must be considered in an interplay of institutional factors - what this work envisages to do with a focus on the educational system. Since better educated young people generally have better employment opportunities the connection to education is theoretically sound. A large number of studies have empirically stated the positive effects for young people of a closer link between training and the labour market through vocational education paths (Allmendinger, 1989; Gangl, 2003; Julkunen, 2010; Kerckhoff, 1995, 2000), which leads to a reduction of youth unemployment (Breen, 2005; Gangl, 2003; Quintini and Manfredi, 2009; Quintini et al., 2007; Wolbers, 2007). Countries that have robust and broad Vocational Education and Training (VET) systems and strong employer-worker linkages had better transitions for labour market entrants so far and have performed best during the recession with regard to labour market outcomes for youth - especially dual apprenticeship systems such as Germany or Austria (Caroleo et al., 2017).

Education systems are characterised by very different features such as investment in education, the teacher-student ratio, compulsory school years, tracking or vocational specificity. From all these facets a factor is sought that represents a close link to the labour market, since the problems of young people (compared to adults) on the labour market are to be investigated. Hence, vocational specificity seems to be the most promising element as it has a major influence on the success rates of the school-to-work transition period (Banerji et al., 2015; Cahuc et al., 2013; Eichhorst et al., 2013; O'higgins, 2001). It is understood as the degree to which an education system sends clear signals to employers about a pupil's qualification and can be measured by the share of students enrolled in vocational tracks. It indicates to what extent vocational specialisation already takes place in schools (Brzinsky-Fay, 2017), whereby it is intended to teach young people practice-oriented knowledge and skills (Eichhorst et al., 2012), which are usually also transferable between different employers (Beck and Katz, 2011). Especially in an age of financial crises and austerity, VET is often seen as a silver bullet for tackling the problem of youth unemployment (Eichhorst et al., 2012). Sociologists have come to the consensus that the two central factors that are important for young people seeking their first job are the degree of specific rather than general skills and the strength of the link between educational systems and the labour market (Allmendinger, 1989; Hannan et al., 1996; Maurice et al., 1986; Müller and Shavit, 1998). When a company wants to hire new employees, who are just leaving school, it is a challenge to assess the potential benefits of this new recruitment as there is no prior work experience to rely on. Thus, one of the main reasons for youth unemployment is the lack of information an employer has about the qualifications of a labour market entrant (Levels et al., 2014; Spence, 1973). A greater focus on vocational training should facilitate transition from school to work by creating a closer link between schools and employers that sends a clear signal to employers about the productivity and skills of a potential applicant (Breen, 2005; Spence, 1973; Stigler, 1962; Stiglitz, 1975).

As important as vocational specificity seems to be for young people and their connection to the labour market, Brzinsky-Fay concludes that this is not the only aspect that can lead to an improvement in youth unemployment, as this area must always be addressed in interaction with other institutions. (Brzinsky-Fay, 2017)

> H4: Vocational Specificity has no main effect on the level of youth unemployment and the youth-to-adult unemployment ratio

In their studies Breen (2005) and Brzinsky-Fay (2017) examine the interplay between vocational specificity and employment protection legislation and show that youth unemployment appears to be low in countries with strong educational signalling and liberal labour market policies. The results imply that an education system with a strong vocational component and a link between education and the labour market can compensate for the negative effects of strong employment protection against dismissal. For the combination of EPL regular and vocational specificity, it is assumed that it is best for young people to have low regulations for regular contracts, so that employers can hire them without fear of high dismissal costs, and, additionally, high educational signalling to ensure early on that employers can assess the abilities of

Table 1: Overview Hypotheses and Previous Results

Notes: Empty fields indicate that these aspects have not yet been investigated in the two highlighted studies, as only EPL on regular contracts has been studied.

Hypotheses	Results from:	
	Breen (2005)	Brzinsky-Fay (2017)
H1: EPL on regular contracts has no significant impact on the level of youth unemployment and the youth-to-adult unemployment ratio	Strict EPL regular = Higher youth-to-adult unemployment	Effect only measured in conjunction with other institutional variables
H2: Strict EPL on temporary contracts is decreasing the level of youth unemployment and the youth-to-adult unemployment ratio	-	-
H3: The level of youth unemployment and the youth-to-adult unemployment ratio is lowest in countries with low EPL regular and high EPL temporary	-	-
H4: Vocational Specificity has no main effect on the level of youth unemployment and the youth-to-adult unemployment ratio	Effect only measured in conjunction with other institutional variables	Effect only measured in conjunction with other institutional variables
H5a: The level of youth unemployment and the youth-to-adult unemployment ratio is lowest in countries with low EPL (regular) and high vocational specificity	confirmed	confirmed
H5b: The level of youth unemployment and the youth-to-adult unemployment ratio is lowest in countries with high EPL (temporary) and high vocational specificity	-	-

potential applicants to establish a strong connection between schools and employers.

> H5a: The level of youth unemployment and the youth-to-adult unemployment ratio is lowest in countries with low EPL (regular) and high vocational specificity

With regard to the interplay between temporary EPL and vocational specificity, it is assumed that young people have the best labour market opportunities if temporary contracts are subject to high regulations so that employers cannot employ them continuously in these insecure and often more poorly paid jobs without regulations.

> H5b: The level of youth unemployment and the youth-to-adult unemployment ratio is lowest in countries with high EPL (temporary) and high vocational specificity

3. Analytical Strategy

3.1. Comparative Research Design

Since youth unemployment levels and the youth-to-adult unemployment ratio as the dependent variables differ substantially across countries and over time, a comparative quantitative analysis with a statistical, variable oriented approach seems most promising (Della Porta, 2008; Lijphart, 1971). Cross-national and time-series data are used to find out how these variables are influenced by the two EPL types and how this relationship may be moderated by vocational specificity. To ensure that the countries are as similar as possible and that sufficiently harmonised data is available to make meaningful comparisons the analysis here limits itself to an OECD sample of 28 countries. As many countries as possible were selected that have different labour market regulations and also represent a balanced sample of Liberal, Conservative, Mediterranean and Social-Democratic employment regimes (Bukodi and Robert, 2007). The case selection is based on variations in the independent variables, i.e. large enough variation when it comes to labour market regulations and vocational approaches in education which is important for a conclusive research design (Della Porta, 2008; King et al., 1994). A timeframe from 1985 - 2013 is selected to cover as many years as possible, including the whole period of EPL data that is currently available from the OECD.

Breen (2005) and Brzinsky-Fay (2017) studies are chosen as the most immediate departure studies as they both examine the influence of EPL on youth labour market chances and its interaction with vocational specificity. While they deal only exclusively with EPL on regular contracts, this paper makes a distinction between EPL on regular and temporary contracts, which hasn't been paid much attention yet (in interactions with vocational specificity). Since young people in particular are increasingly entering the labour market on temporary contracts and deregulation of these contracts in recent years has led to increased uncertainty, a more detailed examination of these two forms can shed new light on the effects of EPL reforms. A study which exclusively examines

the influence of regular EPL on youth in the labour market oversees the majority of young people who are employed on temporary contracts and are not affected by the mechanisms of regular EPL. Compared to both previous studies, the time period is extended by 24 (compared to Breen) and 19 years (compared to Brzinsky-Fay). EPL trends across European and OECD countries have shown fewer changes and reforms regarding regular EPL, but a clear trend towards deregulations of temporary contracts – providing a new contribution by observing both EPL variables over a longer time-period. Lastly, compared to Breen (2005) the defined age group of adults is extended from 25-54 years to 25-64 years, since this division represents the labour market situation much more realistically, based on the average retirement age of 65.8 for men and 65.5 for women across OECD countries (Organisation for Economic Cooperation and Development, 2019).

3.2. Variables and Operationalisation

My dataset comprises observations from 28 OECD countries while covering a time period from 1985 to 2013 to study how EPL for regular and temporary contracts is influencing youth labour market chances and how this link might be conditioned by a country's vocational specificity. Annual data is included from the following countries: Australia (AUS), Austria (AUT), Belgium (BEL), Canada (CAN), Czech Republic (CZE), Denmark (DNK), Finland (FIN), France (FRA), Germany (DEU), Greece (GRC), Hungary (HUN), Ireland (IRL), Italy (ITA), Japan (JPN), Korea (KOR), Mexico (MEX), Netherlands (NLD), New Zealand (NZL), Norway (NOR), Poland (POL), Portugal (PRT), Slovak Republic (SVK), Spain (ESP), Sweden (SWE), Switzerland (CHE), Turkey (TUR), United Kingdom (GBR) and the United States (USA). Summary statistics for each country can be found in the Appendix in Table 1.

3.2.1. Dependent Variables

Two different dependent variables are chosen to reflect the labour market situation of young people: The level of youth unemployment and the youth-to-adult unemployment ratio. The level of youth unemployment is chosen because it was used as the dependent variable in a number of influential previous studies (e.g. Noelke, 2015). Young people are considered to be all people aged 15-24. Even though there is no "clear cut definition of youth" (Eurostat, 2009, p.17), this range is chosen here in order to ensure the greatest possible comparability with previous studies in this research area (e.g. Noelke, Breen and Brzinsky-Fay). Results regarding the interaction of the variables can thus be seen more easily in the context of previous research. Data is taken from the OECD Labour Force Statistics (LFS) database. For the second dependent variable – the ratio of youth to adult unemployment – the same database is used where the youth unemployment rate (15-24 years) is divided by the adult unemployment rate (26-64 years), which is a measure that can partially control for cyclical economic conditions (Breen and Buchmann, 2002). This is in line with research from Breen and Brzinsky-Fay and moreover, in contrast to the youth unemployment rate alone, it can reflect the disadvantages of young people in relation to the adult working group.[2]

3.2.2. Key Explanatory Variables

A central key explanatory variable is the degree of EPL where measures are taken from the OECD Employment database. The strictness of employment protection is divided in two separate measure – indicators on regular and temporary employment – with scores ranging from 0 to 6 and higher values indicating stronger worker protection against dismissal. The 'EPL regular' indicator summarises the inconveniences for employers when they want to dismiss individual workers on regular or open-ended contracts, as well as rules on notice periods, severance pay or valid reasons for dismissal in general. The second indicator 'EPL temporary' refers to the severity of the regulations for the creation of fixed-term contracts and temporary work contracts on the one hand and contractual regulations concerning wages, recruitment procedures and working conditions on the other (Organisation for Economic Cooperation and Development, 2019). Although weaknesses of these indicators were taken into consideration (Eichhorst et al., 2008), they are used here due to a lack of comprehensive alternatives. The period to be examined (1985-2013) was chosen according to the availability of EPL data for the majority of countries. The 1980s show hardly any reform activities without record of any reforms or only changes with a marginal impact (fRDB-IZA Social Reforms Database, 2010). The following 1990s and 2000s, on the other hand, show more significant changes in several countries, both for regular and temporary employment (see Venn, 2009 for a detailed assessment of EPL reforms). My country selection of 28 OECD countries shows six different trends regarding the development of EPL: A partial deregulation, where regular EPL remains constant and only temporary EPL is deregulated, upward and downward trends for both types of EPL, no changes at all, a deregulation of regular contracts combined with a higher regulation of temporary contracts and lastly an initial deregulation with subsequent re-regulation of temporary contracts. Accordingly, Figure 1 in the Appendix shows how the two EPL types have developed from 1985 to 2013 - exemplarily illustrated with individual countries - with blue lines representing regular EPL and orange lines representing temporary EPL.

For the other key explanatory variables – vocational specificity - Worldbank data is used for vocational, general and total enrolment in secondary education. The strength with which an educational system sends clear signals to employers about the performance of potential applicants is calculated by dividing the secondary vocational enrolment by the secondary total enrolment. The selected OECD countries form a large variety with regard to their educational design of the ideal skill formation regimes according to Busemeyer and

[2] Due to data limitations, it is unfortunately not possible to take the NEET rate here, which could potentially capture the labour market situation of young people even better.

Iversen (2012), who classify countries according to the two categories "public investment in VET" and "firm involvement in VET".[3] The literature shows several alternatives as to how different VET systems can be classified and which countries are assigned to them (see as an example Hanushek et al., 2011 in the Appendix Table 2; Eichhorst et al., 2015; Pohl and Walther, 2007). Hadjivassiliou et al. (2015) highlight Germany and the Netherlands as the most successful examples with the most efficient integration of 15-19 years into the labour market with resulting high employment figures among young people over several years and stability during in various economic phases, such as the financial crisis (Hadjivassiliou et al., 2015). Even though some scholars specifically focus on upper secondary education when it comes to the share of vocational education (Breen, 2005) only the general secondary level is used here as the Worldbank provides the most comprehensive dataset regarding my country sample and time frame but does not distinguishing between various levels of secondary education. However, this does not represent a limitation since the expectations do not differentiate between different stages of vocational education nor do other scholars in their research (Caroleo et al., 2017). It should be further noted at this point that I am aware that this indicator represents a questionable operationalisation, since no distinction can be made between school-based and workplace-based training and the pure enrolment statistics can be misleading without sufficient background knowledge of the design of the respective education systems of the countries. It is precisely a dual system of apprenticeship, as in the German-speaking countries, that can ensure a smooth school-to-work transition through a positive effect of vocational specificity (Russell and O'Connell, 2001; Van Der Velden and Wolbers, 2003). However, there are countries such as Italy or the Czech Republic, which also have high enrolment rates in vocational education programmes, but their link between schools and employers is very weak and thus a lack of specific skills of vocational pupils does not result in better positioning on the labour market (Avdagic, 2015). But since no comprehensive data is yet available for these detailed differences, this analysis draws on the general share of vocational enrolment, whereby a closer examination of these differences can be an exciting task for future research.

3.2.3. Control Variables

As it is the core of statistical interference to incorporate control variables in order to consider potential confounding factors (Angrist and Pischke, 2009), three macroeconomic controls are used in this study. Poor macroeconomic conditions can increase competition for jobs between young people, making smooth transitions into the labour market more difficult and exposing young people to the risk of dismissal due to the overall economic situation, as they are the least protected (Gangl, 2002). GDP per capita from the Worldbank database is used to control for the impact of macroeconomic conditions on activities of the welfare state and public expenditures for labour market programmes. In order to capture the influence of inequality, the GINI coefficient from the World Inequality Database (WID) and the Estimated Household Income Inequality Data Set (EHII) is also taken into account. Finally, the general level of unemployment is controlled for with data on general unemployment (15+ years) from the International Labour Organisation (ILO) but only in models where the ratio of youth-to-adult unemployment is used as a dependent variable. All data used was collected before this research and is publicly available.

3.3. Methods

To find out what influence the two versions of EPL have on the level of youth unemployment and the youth-to-adult unemployment ratio and whether this relationship may be moderated by vocational specificity, OLS regressions with robust standard errors are calculated. As the individual observations within a country are not independent, the standard errors are clustered by countries to control for the correlation between the errors within the same country-unit over time. This technique is used to generate robust estimates of standard errors that compensate for the fact that the observations are independent between countries, but not within them (see also Breen, 2005).

3.3.1. Missing Data

The biggest challenge in creating the dataset is the problem of missing data points. This missingness was addressed by using linear interpolation to fill in missing values, assuming a country-specific linear time trend for the gaps in the data (StataCorp, 2015). A maximum of four missing years were interpolated. The fact that a variable has no values at the beginning or at the end of the sample period was not addressed further, as no substantial statistical problems arise (Beck and Katz, 2011). Missing data should ideally be addressed by multiple imputation (Beck and Katz, 2011), where relevant information is extracted from the data and then multiple values are imputed for each data-gap, creating multiple "complete" datasets (Honaker and King, 2010; Honaker et al., 2011). Even though this does not limit the following analyses, it can be addressed in future research.

3.3.2. Robustness Checks

To ensure that my results are not driven by any one country, I re-estimate all models excluding one country at a time. The evaluation of these models can provide information on whether the model is susceptible to changes in the sample and whether individual countries strongly influence the overall results. Furthermore, a distinction of two periods is made (pre-2000 and post-2000) to check whether the results differ based on the time period chosen. This is to control for major institutional changes over time.

[3] For a very detailed overview of an analysis of different country clusters regarding vocational education see Zimmermann et al., 2013.

3.3.3. Stages of Analysis

At the beginning, all variables are displayed descriptively first. For this purpose, an average is calculated for all 28 countries for each year from 1985 to 2013 in order to make general trends visible across all countries. The results can be seen in Figures 1-4. Then OLS regressions with robust standard errors are calculated for different models. All analyses are calculated with the level of youth unemployment as the first dependent variable (Table 1), followed by the ratio of youth-to-adult unemployment as the second dependent variable (Table 2). A total of nine models are calculated each for both dependent variables. Models 1-3 test the influence of the two types of EPL on the dependent variable. Control variables are not included in any model. Model 1 tests the sole influence of EPL for regular contracts, Model 2 the sole influence of EPL for temporary contracts and Model 3 combines both types. Starting with Model 4, the two control variables GDP per capita and GINI coefficient are added to the level of youth unemployment, while the overall unemployment rate is added additionally in Table 2. In Model 5, the third key explanatory variable vocational specificity is introduced, so that in this model all key explanatory variables are examined together. Finally, the Models 6-9 test interaction effects, whereby Model 6 examines the interaction between EPL regular and EPL temporary, Model 7 the interaction between EPL regular and vocational specificity, Model 8 the interaction between EPL temporary and vocational specificity and Model 9 all three interaction terms. Robustness checks are displayed in Tables 3 to 6.

4. Results

4.1. Descriptive Results

Figure 1 compares the development of youth unemployment rates and adult unemployment rates across all 28 countries from 1985 to 2013. It shows that the level of adult unemployment (orange line) has hardly changed over the entire period and moves exclusively between the 5 and 10% mark. Youth unemployment (blue line), however, shows strong fluctuations and, with one exception in 2010, is always higher than adult unemployment, so young people seem to have it more difficult in the labour market than adults when it comes to their employment prospects. The figures for youth unemployment range from around 7% (in 2010) to peak values of over 30% in the crisis years 2007/2008, showing a significant negative impact of the financial crisis on the labour market situation of young people (compared to a moderate impact on their adult counterparts). Figure 2 adopts the blue graph of youth unemployment and now shows in orange - displayed on the secondary Y-axis - the youth-to-adult unemployment rate, which lies predominantly between 2.5 and 3%. Hence, Figure 2 shows the two dependent variables to be examined. Figures 3 and 4 show the three key explanatory variables, whereby EPL for regular and EPL for temporary contracts are summarized together in Figure 3. Across all countries temporary EPL saw a clear downward trend and is continuously deregulated, while regular EPL remains mostly constant. Furthermore, it can be seen that already at the beginning in 1985 regular contracts are regulated much more strictly than temporary contracts, whereby regulations for regular contracts at the end of the measurement period in 2013 just reach the starting point of regulation for temporary contracts with a value of 2 (on a scale of 0-6). The often-stated trend towards 'partial deregulation' is thus also evident in the countries used here. Examples of individual countries showing various EPL trends over time can be found in Figure 1 (Appendix). Figure 4 concludes by summarising the development of vocational specificity in secondary education, although no clear trend can be discerned here. Overall, the values are between 22 and 30%, which describes the proportion of pupils who are in a vocational track compared to the total enrolment in secondary education across all countries. A more detailed country-by-country table displaying the main descriptive results of the dependent and all independent variables can be found in the Appendix (Table 1). Figure 5 plots youth unemployment and the youth-to-adult unemployment rate against both EPL indicators for regular and temporary contracts. In this comprehensive sample there appears to be little support for the often-proclaimed relationship between strict regulation and higher youth unemployment when looking at the bivariate associations.

4.2. Regression Analyses

Table 1 reports estimates from ordinary least squares (OLS) regression with country clustered robust standard errors using the level of youth unemployment. Table 2 adopts the same model building structure but uses the youth-to-adult unemployment ratio instead.

In Table 1 no regression coefficient of the main effects or the interaction terms is statistically significant. It shows that neither EPL for regular contracts, nor EPL for temporary contracts, nor the vocational specificity have a significant influence on the level of youth unemployment in the 28 countries of this sample. Table 2 shows the same nine models but this time using the ratio of youth-to-adult unemployment as the dependent variable. Again, a stronger regulation of regular contracts has no statistically significant impact on the ratio of youth-to-adult unemployment. The results for both dependent variables thus confirm the results of Noelke (2015) who found no consistent evidence that stronger regulation of regular contracts is associated with higher youth unemployment. Yet they disagree with Breen (2005) who found a significant effect indicating that higher EPL for regular contracts will lead to an increase in youth unemployment.

In Table 2 the coefficients of EPL temporary are statistically significant at the 5% level in Models 2 to 5, showing a positive coefficient. Temporary EPL does therefore not affect the level of youth unemployment but the ratio of youth-to-adult unemployment. This means that in countries with a stricter regulation of temporary contracts, the number of unemployed youth increases compared to unemployed adults so young people are more affected by negative consequences

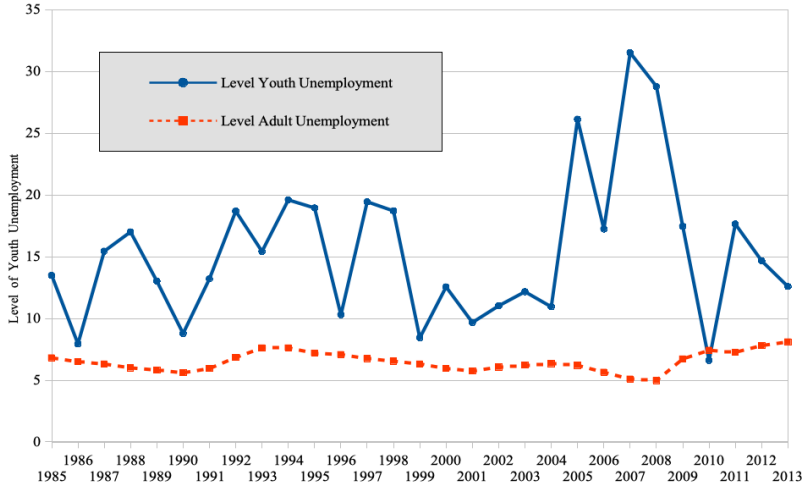


Figure 1: Levels of youth unemployment and levels of adult unemployment (average across 28 OECD countries), 1985-2013

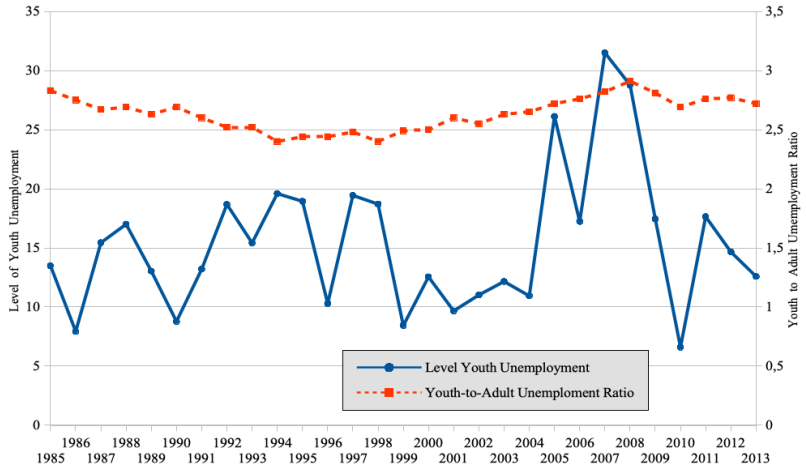


Figure 2: Levels of youth unemployment and the youth-to-adult unemployment ratio (average across 28 OECD countries), 1985-2013

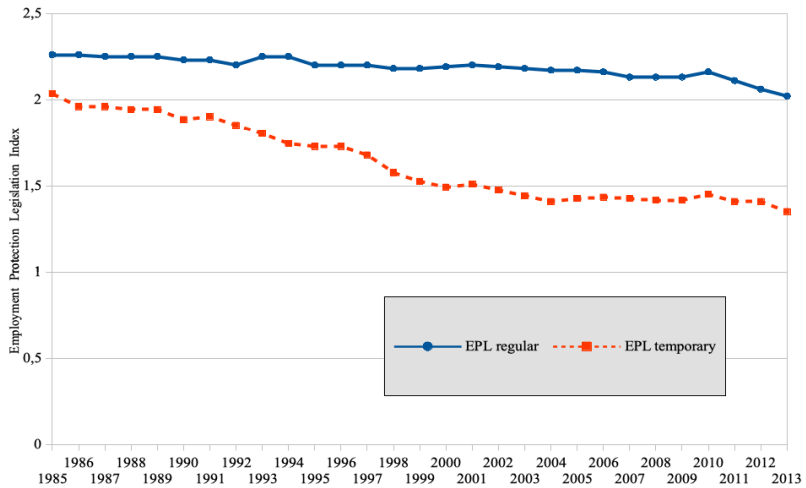


Figure 3: Employment protection legislation for regular and temporary contracts (average across 28 OECD countries), 1985-2013

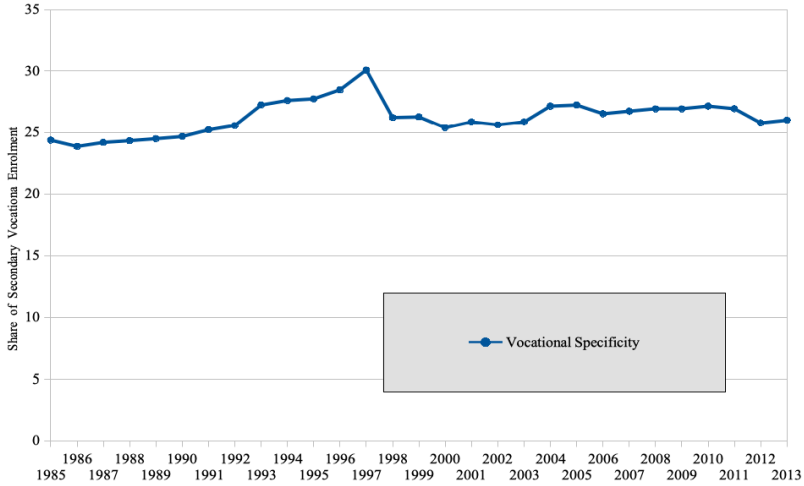


Figure 4: Share of secondary vocational enrolment in % (average across 28 OECD countries), 1985-2013

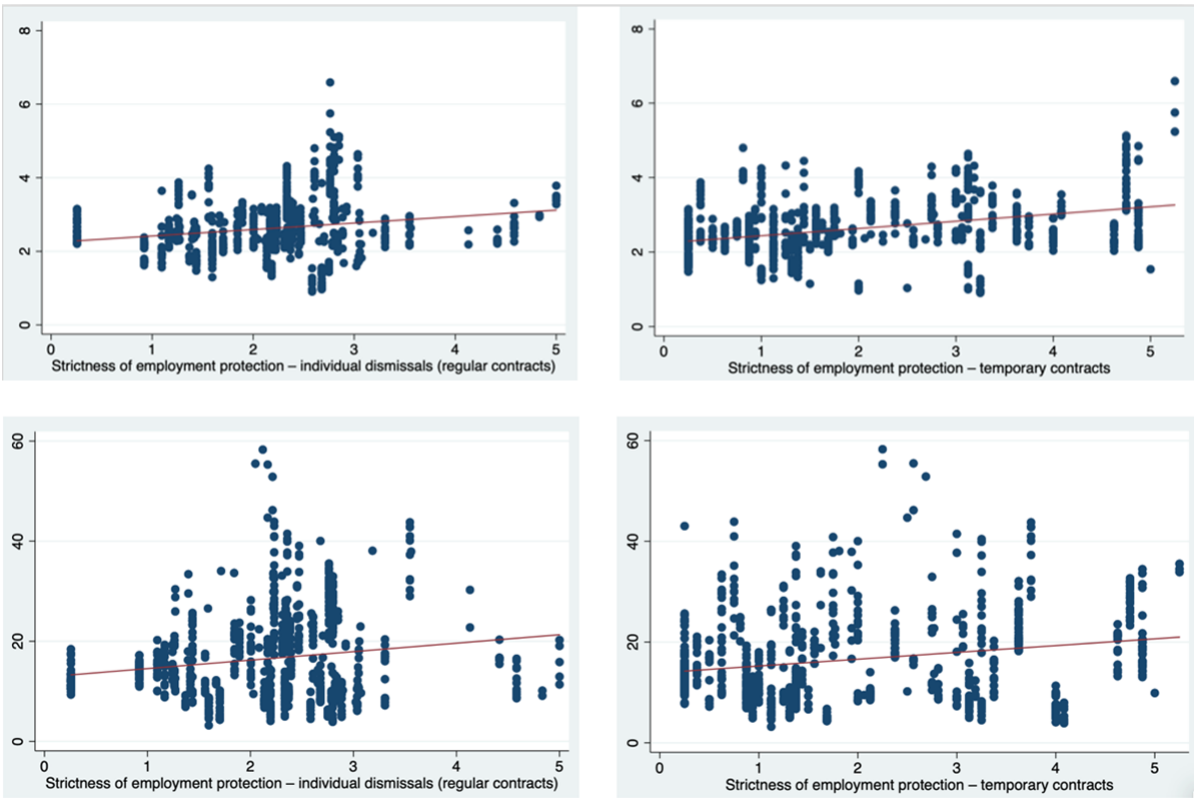


Figure 5: Bivariate associations between both types of EPL and youth-to-adult unemployment ratio (first row) and youth unemployment rates (second row), 28 OECD countries, 1985–2013

regarding their employment status than adults in the same country. Breen and Brzinsky-Fay did not investigate this relationship, but the results contradict the in previous literature embedded Hypothesis 2 that higher regulations of temporary EPL have a positive effect on young people and their labour market chances. Moreover, the interaction term of EPL regular and EPL temporary in Table 2 is significant in models 6 and 9. These positive coefficients states that higher levels of EPL regular strengthen the positive effect of EPL temporary. In a country with strict EPL on temporary contracts the labour market situation for young people is therefore worse compared to adults and they are more often unemployed, and this trend will be even stronger in countries that also have strict EPL on regular contracts. Vocational Specificity is not significant in any of the models as a main effect, confirming Hypothesis 4. Model 9 including all three interactions achieves the highest model fit (R-squared = 0.2557) whereby approx. 1/4 of the variance of the dependent variable can be

Table 2: OLS regression with country clustered robust standard errors – Level of youth unemployment

Country clustered robust standard errors are reported in parentheses. The constant is not reported. *** Significant at the 1% level. ** Significant at the 5% level. * Significant at the 10% level.

Dependent variable Level of youth unemployment	M1	M2	M3	M4	M5	M6	M7	M8	M9
Key Explanatory Variables									
EPL Regular	1.69 (1.16)	-	0.73 (1.02)	1.18 (1.25)	0.31 (1.30)	-2.73 (2.38)	-0.06 (1.91)	1.61 (1.86)	-0.81 (2.72)
EPL Temporary	-	1.35 (0.85)	1.13 (0.92)	0.48 (1.00)	0.19 (1.06)	-3.84 (3.37)	0.22 (1.10)	-2.35 (2.09)	-4.80 (3.75)
Vocational Specificity	-	-	-	-	0.12 (0.09)	0.19* (0.10)	0.07 (0.22)	-0.04 (0.15)	0.06 (0.233)
Interaction Effects									
EPL regular* EPL temporary	-	-	-	-	-	1.73 (1.37)	-	-	1.29 (1.13)
EPL regular* vocational specificity	-	-	-	-	-	-	0.02 (0.10)	-	-0.01 (0.11)
EPL temporary* vocational specificity	-	-	-	-	-	-	-	0.09 (0.07)	0.07 (0.06)
Control Variables									
GDP per capita	-	-	-	0.00** (0.00)	0.00** (0.00)	0.00** (0.00)	0.00** (0.00)	0.00** (0.00)	0.00** (0.00)
GINI coefficient	-	-	-	39.95 (29.97)	57.92* (33.78)	59.38* (31.74)	58.78 (34.84)	68.61** (29.48)	67.09** (29.82)
N	744	744	744	717	627	627	627	627	627
R2	0.0251	0.0456	0.0491	0.1377	0.1515	0.1707	0.152	0.1719	0.1816

explained by this model.

Tables 3-6 show a robustness check with two separate time-periods – before and after 2000. Again, both dependent variables are examined - the level of youth unemployment in Tables 3 and 4 and the youth-to-adult unemployment ration in Tables 5 and 6. Seven different models are calculated for each time period. Model 1 examines the influence of the two EPL key explanatory variables (EPL regular and EPL temporary) on the dependent variable. Model 2 adds vocational specificity. Models 3 to 5 additionally add single interactions while Models 6 and 7 combine interactions. All models include all control variables.

The robustness check with the level of youth unemployment as a dependent variable (Tables 3 and 4) shows a significant interaction of the two types of EPL before 2000 as already interpreted above. Moreover, M5 in Table 3 reveals a slightly significant value at the 10% level for the interaction of EPL temporary and vocational specificity. No coefficient in the after 2000 period is significant. Tables 5 and 6 with the ratio of youth-to-adult unemployment show several statistically significant values. Again - before 2000 - the EPL regular EPL temporary interaction is significant. Furthermore, it appears that stricter temporary EPL increases youth unemployment significantly more than adult unemployment. This is particularly the case before 2000 (Table 5). After 2000, this link appears to be conditioned by the degree of vocational specificity implying that the association only shows up in countries with higher level of vocational specificity (M 5-7 Table 6). The positive effect of temporary EPL on the youth-to-adult unemployment rate is therefore strengthened by a higher level of vocational specificity – and not weakened as assumed in theoretical section where it is argued that a stronger vocational orientation sends clear signals to employers about the skills of young people and thus reduces youth unemployment. So far, no sound arguments for this connection can be found in the literature. It is assumed that vocational specificity often only works in conjunction with other factors, but even then, only positively for young people. It might thus be the case that the very small effects with coefficients around 0.01 and a marginal significance at the 10% level may no longer be visible with a different country or time sample and varying variable operationalisations. As a further robustness check, the regressions were recalculated, whereby each time a different country was excluded from the analysis in order to test whether the results are driven by one country. However, the results show that the model is not susceptible to changes in the sample (see Tables 3-6 in the Appendix).

Breen's findings on the interaction between regular EPL and vocational specificity show that vocational specificity can compensate for the problem of higher youth unemployment in countries with strict EPL regulation. It is confirmed that the ratio of youth-to-adult unemployment is lowest in countries with high educational signalling (Breen, 2005).

Table 3: OLS regression with country clustered robust standard errors – Ratio of youth-to-adult unemployment

Country clustered robust standard errors are reported in parentheses. *** Significant at the 1% level. ** Significant at the 5% level. * Significant at the 10% level.

Dependent variable Ratio of youth-to-adult unemployment	M1	M2	M3	M4	M5	M6	M7	M8	M9
Key Explanatory Variables									
EPL Regular	0.18 (0.11)	-	0.01 (0.08)	0.04 (0.08)	0.05 (0.11)	-0.38* (0.22)	-0.02 (0.16)	0.12 (0.17)	-0.44 (0.34)
EPL Temporary	-	0.19** (0.08)	0.19** (0.08)	0.20** (0.09)	0.22** (0.10)	-0.38 (0.26)	0.22** (0.10)	0.08 (0.24)	-0.36 (0.28)
Vocational Specificity	-	-	-	-	0.00 (0.01)	0.01 (0.01)	-0.01 (0.02)	-0.01 (0.01)	0.00 (0.02)
Interaction Effects									
EPL regular* EPL temporary	-	-	-	-	-	0.25** (0.11)	-	-	0.25** (0.11)
EPL regular* vocational specificity	-	-	-	-	-	-	0.00 (0.01)	-	0.00 (0.01)
EPL temporary* vocational specificity	-	-	-	-	-	-	-	0.00 (0.01)	0.00 (0.01)
Control Variables									
GDP per capita	-	-	-	5.17e-06 (4.12e-06)	4.29e-06 (4.66e-06)	5.00e-06 (4.96e-06)	5.41e-06 (4.5.4e-06)	4.96e-06 (4.69e-06)	5.44e-06 (4.74e-06)
GINI coefficient	-	-	-	1.26 (2.78)	1.31 (2.87)	1.62 (2.60)	1.48 (2.91)	1.94 (2.62)	1.82 (2.51)
Overall Unemployment				-0.03 (0.02)	-0.03 (0.02)	-0.03* (0.02)	-0.03 (0.02)	0.03** (0.02)	-0.03* (0.02)
N	744	744	744	717	627	627	627	627	627
R2	0.0363	0.1268	0.127	0.1737	0.2008	0.2537	0.2030	0.2085	0.2557

Brzinsky-Fay corroborates this connection using QCA, confirming that strict regular EPL is only associated with high relative youth unemployment in case of low vocational specificity. However, vocational training alone is not sufficient to combat youth unemployment, as this must always be seen in conjunction with other institutions (Brzinsky-Fay, 2017).

As the findings of this thesis do not coincide with the results of Breen (2005) and Brzinsky-Fay (2017), individual countries will be examined in more detail in order to possibly get a better understanding of the results. Concrete country examples will be used to show how countries can be categorised with regard to regular and temporary EPL and how the two dependent variables look in the respective categories. For all 28 countries, the mean value for regular and temporary EPL over the entire period is used. Then the median for both types of EPL is calculated, whereby all countries above the median are divided into "strong" EPL and all below the median into "weak" EPL. The mean value of the dependent variables for each country is also calculated. Tables 5 and 6 show the allocation of countries based on the EPL categories and lists all countries if their level of youth unemployment or their youth-to-adult unemployment rates are above the median. Here, too, a division into before and after 2000 has been made.

The first table showing the level of youth unemployment finds no clear relationship between EPL regulations and youth unemployment. However, Table 8, where the ratio of youth-to-adult unemployment was chosen, shows a pattern regarding the relationship between EPL and the youth-to-adult unemployment ratio. In line with the regression analyses, countries where the youth are more hit by unemployment than the adults primarily tend to be those with strong EPL regulations, especially strong temporary EPL regulations. This country overview shows that significantly more countries have a comparatively high youth-to-adult unemployment rate if both temporary and regular contracts are subject to high regulations (pre-2000 in seven countries, post-2000 in six countries), while only two countries with weak EPL on regular and strong EPL on temporary contracts have comparatively worse labour market chances for young people compared to adults. Thus, in a country with strong EPL on regular contracts and also strong EPL on temporary contracts, the labour market situation of young people is worse compared to that of adults, which is the case for example in France, Greece or Italy both before and after 2000.

Table 10 shows once again all hypotheses and the results

Table 4: Robustness Check before 2000 – Level of youth unemployment

Dependent variable	Before 2000						
Level of youth unemployment	M1	M2	M3	M4	M5	M6	M7
Key Explanatory Variables							
EPL Regular	0.14	-0.62	-8.20**	-1.41	2.28	2.41	-5.17
	(1.78)	(1.91)	(3.29)	(1.83)	(2.71)	(2.96)	(4.30)
EPL Temporary	1.05	1.02	-7.81**	1.09	-3.35	-3.40	-8.96**
	(1.21)	(1.20)	(3.35)	(1.23)	(2.85)	(3.03)	(3.66)
Vocational Specificity	-	0.12	0.33**	-0.01	-0.13	-0.11	0.18
		(0.08)	(0.11)	(0.20)	(0.15)	(0.20)	(0.18)
Interaction Effects							
EPL regular* EPL temporary	-	-	3.84**	-	-	-	3.23**
			(1.44)				(1.29)
EPL regular* vocational specificity	-	-	-	0.06	-	-0.01	-0.01
				(0.09)		(0.08)	(0.09)
EPL temporary* vocational specificity	-	-	-	-	0.14*	0.14	0.08
					(0.08)	(0.08)	(0.06)
Control Variables							
GDP per capita	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
GINI coefficient	18.78	25.87	23.01	27.79	51.41	51.43	38.05
	(41.35)	(45.61)	(41.65)	(46.86)	(44.30)	(44.23)	(42.95)
N	333	308	308	308	308	308	308
R2	0.1811	0.2023	0.3231	0.2072	0.2686	0.2687	0.3417

Table 5: Robustness Check after 2000 – Level of youth unemployment

Dependent variable	After 2000						
Level of youth unemployment	M1	M2	M3	M4	M5	M6	M7
Key Explanatory Variables							
EPL Regular	0.71	-0.74	-0.38	-1.65	-0.55	-1.68	-1.30
	(1.19)	(1.39)	(4.51)	(3.63)	(1.58)	(3.59)	(6.12)
EPL Temporary	0.75	-0.07	0.38	0.00	-0.82	-0.89	-0.56
	(1.05)	(1.14)	(5.39)	(1.18)	(2.59)	(2.58)	(5.92)
Vocational Specificity	-	0.09	0.09	0.00	0.03	-0.10	-0.10
		(0.09)	(0.12)	(0.36)	(0.24)	(0.43)	(0.43)
Interaction Effects							
EPL regular* EPL temporary	-	-	-0.20	-	-	-	-0.17
			(2.41)				(2.38)
EPL regular* vocational specificity	-	-	-	0.04	-	0.05	0.05
				(0.17)		(0.16)	(0.17)
EPL temporary* vocational specificity	-	-	-	-	0.03	0.04	0.04
					(0.11)	(0.11)	(0.11)
Control Variables							
GDP per capita	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
GINI coefficient	19.27	45.96	46.77	47.20	45.30	46.77	47.27
	(32.64)	(36.64)	(42.33)	(39.89)	(37.43)	(39.80)	(42.77)
N	384	319	319	319	319	319	319
R2	0.1441	0.1546	0.1547	0.1553	0.1556	0.1567	0.1568

Country clustered robust standard errors are reported in parentheses.
*** Significant at the 1% level. ** Significant at the 5% level. * Significant at the 10% level.

Table 6: Robustness Check before 2000 – Ratio of youth-to-adult unemployment

Country clustered robust standard errors are reported in parentheses.
*** Significant at the 1% level. ** Significant at the 5% level. * Significant at the 10% level.

Dependent variable	Before 2000						
Level of youth unemployment	M1	M2	M3	M4	M5	M6	M7
Key Explanatory Variables							
EPL Regular	-0.02	-0.04	-0.76**	-0.13	0.06	-0.04	-0.90**
	(0.13)	(0.15)	(0.29)	(0.18)	(0.20)	(0.22)	(0.37)
EPL Temporary	0.29**	0.36**	-0.49*	0.36**	0.21	0.25	-0.42
	(0.13)	(0.12)	(0.28)	(0.12)	(0.31)	(0.31)	(0.34)
Vocational Specificity	-	-0.00	0.02*	-0.02	-0.01	-0.02	0.01
		(0.01)	(0.01)	(0.02)	(0.01)	(0.02)	(0.01)
Interaction Effects							
EPL regular* EPL temporary	-	-	0.37**	-	-	-	0.37**
			(0.12)				(0.12)
EPL regular* vocational specificity	-	-	-	0.01	-	0.01	0.00
				(0.01)		(0.01)	(0.01)
EPL temporary* vocational specificity	-	-	-	-	0.00	0.00	0.00
					(0.01)	(0.00)	(0.01)
Control Variables							
GDP per capita	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**	0.00**
	(7.78e-06)	(9.73e-06)	(0.00)	(0.00)	(9.81e-06)	(0.00)	(0.00)
GINI coefficient	1.47	-1.20	-1.21	-0.96	-0.27	-0.31	-1.57
	(4.07)	(3.91)	(3.43)	(3.91)	(3.72)	(3.70)	(3.51)
Overall Unemployment	-0.04*	-0.04*	-0.06**	-0.04*	-0.05**	-0.05**	-0.06**
	(0.02)	(0.21)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
N	333	308	308	308	308	308	308
R2	0.3618	0.4263	0.5249	0.4328	0.4326	0.4361	0.5285

of this thesis regarding the two dependent variables. Only the results over the entire period from 1985 – 2013 are reported in the table, while further specifications from the robustness checks before and after 2000 are noted with asterisks if they differ from the general results. This overview suggests a complex interplay of regular and temporary EPL, and it seems conceivable that the temporary EPL effect is somehow conditioned on the level of regular EPL so these two do not act independently. A possible explanation for these unexpected results that temporary EPL has a positive and not a negative effect on the youth-to-adult unemployment rate is, that in case of strict temporary EPL employers are reluctant to hire young people even for insecure 'labour market fringe' jobs, as it is difficult to dismiss them later. The reasoning would therefore be similar to that of strict EPL for regular contracts, where it is assumed that employers would rather not hire any new employees than those for whom they later fear high dismissal costs. It would be possible that the nature of the employment contract is not a determining factor for employers to decide whether or not to recruit. If the regulations and thus the costs associated with dismissal are high, fewer people will be hired on both, regular and temporary contracts. Since young people enter the labour market primarily via temporary contracts, excessively high regulations can lead to a closure of this gate to the labour market for them, since employers do not hire new people. According to these results, deregulation of temporary contracts has a positive effect on the labour market situation of young people. For adults, however, the assumed best case is a combination of high EPL for regular contracts, as this protects employers as labour market insiders from dismissals, but also high regulations on temporary contracts, as employers would otherwise have a huge incentive to simply convert regular contracts into more insecure and less paid temporary contracts.

Gebel and Giesecke (2016) show two outcomes resulting from a partial deregulation: First, employers are expected to hire more workers because temporary contracts exonerate them from the otherwise feared dismissal costs. This results in a 'buffer' of temporary jobs which employers can build up and which can be quickly stocked up or dismantled depending on the market situation (Polavieja, 2003). For young labour market entrants, deregulation therefore means that jobs are created for them that would not exist otherwise with a higher regulation of temporary contracts. Secondly, it is argued that employers use temporary contracts as a 'screening tool' (Korpi and Levin, 2001) and thus test whether the worker's skills match the company's needs prior to final permanent recruitment. This mechanism is particularly relevant for young people, as employers have no information about the productivity in the labour market of new labour market

Table 7: Robustness Check after 2000 – Ratio of youth-to-adult unemployment

Country clustered robust standard errors are reported in parentheses.
*** Significant at the 1% level. ** Significant at the 5% level. * Significant at the 10% level.

Dependent variable	After 2000						
Level of youth unemployment	M1	M2	M3	M4	M5	M6	M7
Key Explanatory Variables							
EPL Regular	-0.06	-0.08	-0.45	-0.37	0.00	-0.38	-0.99
	(0.12)	(0.15)	(0.41)	(0.31)	(0.18)	(0.29)	(0.77)
EPL Temporary	0.15	0.09	-0.37	0.11	-0.22	-0.24	-0.75
	(0.09)	(0.09)	(0.40)	(0.11)	(0.19)	(0.18)	(0.46)
Vocational Specificity	-	0.00	0.01	-0.02	-0.02	-0.06*	-0.07*
		(0.01)	(0.01)	(0.03)	(0.02)	(0.03)	(0.04)
Interaction Effects							
EPL regular* EPL temporary	-	-	0.21	-	-	-	0.26
			(0.19)				(0.23)
EPL regular* vocational specificity	-	-	-	0.01	-	0.02	0.02
				(0.01)		(0.01)	(0.01)
EPL temporary* vocational specificity	-	-	-	-	0.01*	0.02**	0.01*
					(0.01)	(0.01)	(0.01)
Control Variables							
GDP per capita	3.47e-06	4.33e-06	3.02e-06	4.86e-06	3.34e-06	3.90e-06	2.72e-06
	(4.32e-06)	(4.74e-06)	(5.40e-06)	(4.40e-06)	(4.48e-06)	(4.03e-06)	(4.74e-06)
GINI coefficient	-3.02	-0.83	-1.71	-0.44	-1.15	-0.67	-1.50
	(2.18)	(2.15)	(2.09)	(2.30)	(2.15)	(2.12)	(2.03)
Overall Unemployment	-0..01	-0.02	-0.01	0.02	-0.01	-0.01	-0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
N	384	319	319	319	319	319	319
R2	0.0788	0.0657	0.0897	0.0803	0.1024	0.1289	0.1600

Table 8: Country Examples – Level of youth unemployment

Dependent Variable Level of youth unemployment		Countries with youth-to-adult unemployment rates >median	
EPL regular	EPL temporary	Before 2000	After 2000
Strong	Strong	FRA, GRC, ITA, ESP, TUR	FRA, GRC, ITA, POL, PRT, ESP, TUR
	Weak	FIN, SVK	CZE, SVK, SWE
Weak	Strong	BEL	BEL, FIN
	Weak	AUS, CAN, HUN, IRL, POL, GBR	HUN, IRL

Table 9: Country Examples – Youth-to-adult unemployment rate

Dependent Variable Youth-to-adult unemployment ratio		Countries with youth-to-adult unemployment rates >median	
EPL regular	EPL temporary	Before 2000	After 2000
Strong	Strong	FRA, GRC, ITA, KOR, PRT, SWE, TUR	FRA, GRC, ITA, KOR, NOR, POL
	Weak	FIN, SVK	CZE, SWE
Weak	Strong	BEL, NOR	BEL, FIN
	Weak	USA, NZL, POL	HUN, NZL, GBR

Table 10: Overview Hypotheses and Results (1985-2013)

Notes: For some hypotheses the robustness checks reveal different results before and after 2000
* Strict EPL (temporary) = High youth-to-adult unemployment. ** Before 2000: Strict EPL (regular and temporary) = High youth-to-adult unemployment
*** Before 2000: Strict EPL (regular and temporary) = High youth-to-adult unemployment, After 2000: No interaction effect
**** After 2000: Strict EPL (temporary) and high vocational specificity = High youth-to-adult unemployment

Hypotheses	Level of youth unemployment	Youth-to-adult unemployment ratio
H1: EPL on regular contracts has no significant impact on the level of youth unemployment and the youth-to-adult unemployment ratio	Confirmed	Confirmed
H2: Strict EPL on temporary contracts is decreasing the level of youth unemployment and the youth-to-adult unemployment ratio	Disproved - No main effect	Disproved*
H3: The level of youth unemployment and the youth-to-adult unemployment ratio is lowest in countries with low EPL regular, high EPL temporary	Disproved** - No interaction effect	Disproved***
H4: Vocational Specificity has no main effect on the level of youth unemployment and the youth-to-adult unemployment ratio	Confirmed	Confirmed
H5a: The level of youth unemployment and the youth-to-adult unemployment ratio is lowest in countries with low EPL (regular) and high vocational specificity	Disproved - No interaction effect	Disproved - No interaction effect
H5b: The level of youth unemployment and the youth-to-adult unemployment ratio is lowest in countries with high EPL (temporary) and high vocational specificity	Disproved - No interaction effect	Disproved ****

entrants. More often they do not get a chance at all from employers to show their skills on the labour market in a highly regulated environment. This also shows that a stronger regulation of temporary contracts and the elimination of this screening tool for employers has different effects for young people and adults. Adults will most likely already have a labour market history, so that employers can assess their productivity and fit for the company by their previous employments when hiring new adults. For young people, however, there is no already existing work history in their school-to-work transition, so that a stronger regulation of temporary contracts leads to fewer young people being hired for new jobs compared to adults.

After a possible explanation was given why a strict regulation of temporary contracts has a more severe effect on young people and why a deregulation of these contracts can be positive for the labour market situation of young people in comparison to adults, a brief look will be taken at why this thesis possibly came to different results than Breen (2005) or Brzinsky-Fay (2017). The first difference is the choice of countries and the period to be examined. Breen only examined a period from 1995-1999, so his results are more of a snapshot without being able to pick up on changes and trends over time. Brzinsky-Fay takes a period from 1990 to 2009, which expands Breen's study, but covers fewer years than this thesis from 1985 to 2013. The end is set for 2009 to avoid possible distortions from the global economic crisis that might alter the post-2000 results of this thesis. Furthermore, it is the novelty of this thesis that for the first time the interplay between the two EPL forms (regular and temporary) and vocational specificity was examined. In addition, different results can always be obtained if different sources were used and the core variables were operationalised differently. Vocational specificity by Breen was confined to upper secondary education and in Brzinsky-Fay not further specified, while in this thesis the vocational enrolment was chosen for the entire secondary education. Breen and Brzinsky-Fay use the OECD Education at a Glance publications for this purpose, while this thesis vocational enrolment was manually calculated using the OECD database.

5. Conclusion

The results presented here support the argument from Noelke (2015) that there is no effect of regular EPL on the youth unemployment rate alone. High regulations on regular contracts only are thus not to blame for a high-level youth unemployment or high youth-to-adult unemployment rates. In line with previous research, it was expected that the deregulation of temporary contracts would have a negative impact on the labour market situation of young people, as they would be less protected. However, the results of this thesis show a different picture - namely that a stronger regulation of temporary contracts has a significantly negative effect on the youth-to-adult unemployment rate. One possible explanation for this is that employers are more inhibited in hir-

ing new employees under higher regulation - on both regular and temporary contracts - because they want to avoid higher dismissal costs. Since young people are particularly often recruited via temporary contracts and the screening function of these contracts is an important indicator of the productivity of future employees, stronger regulations of these contracts have negative effects especially for young people, but not so much for adults. The conclusion that can be drawn here is that the effect of EPL for temporary contracts may not be as one-sided as previously assumed in the literature, and that - as with EPL for regular contracts - there are ambiguous results on its effect, depending on how a model is specified and which countries and groups were considered over which period.

As this thesis was built upon the work of Breen (2005) and Brzinsky-Fay (2017) it was expected to find a statistically significant interaction between EPL and the vocational specificity of an educational system. However, these findings could not be validated, as no interaction was significant over the whole time-period from 1985-2013. The robustness check splitting this time in a pre- and post-2000 period however revealed, that at least for the ratio of youth-to-adult unemployment the post-2000 period showed a significant interaction of EPL for temporary contracts and vocational specificity. The interaction though implies a negative effect of high vocational specificity for young people, so that due to a lack of sound theoretical arguments for this it is assumed that the very small effects, which are predominantly only significant at the 10% level, might be eliminated by changes in the country sample or the time period. Despite these seemingly ambiguous results, it should however be clearly said that when designing policies to reduce youth unemployment, institutions should not be considered as stand-alone variables, but always in their complex interaction with other institutions.

For future studies, it should be noted that certainly other factors besides employment protection legislation and vocational specificity influence youth unemployment, such as union strength and bargaining power, which should be investigated. Micro level and longitudinal data could make it possible to include more control variables and test more detailed relationships and hypotheses. Furthermore, it has already been pointed out in this thesis that the classification of educational signalling based on the enrolment rate does not adequately reflect the actual education systems of the individual countries. Here, more in-depth country analysis, both qualitative and quantitative, could provide a better overview. Furthermore, it is possible to look at the labour market situation of young people not only by looking at the distinction between employed and unemployed but also by looking at other indicators that can capture the situation of young people (in comparison to adults), such as temporary employment rates. An examination of the NEET rate instead of the unemployment rates can moreover provide detailed insights and a more realistic picture of the youth labour market if sufficient data is available. Lastly, in this thesis, groups were formed exclusively on the basis of age, while the influence of labour market regulations possibly also has different consequences depending on one's educational level and for men and women. Future research should also see whether the results can be generalised here, preferably by including non-OECD countries.

References

Addison, J. and Teixeira, P. The economics of employment protection. *Journal of Labor Research*, 24:85–127, 2003.

Allard, G. and Lindert, P. H. Euro-productivity and euro-jobs since the 1960s: Which institutions really mattered? (No. w12460). National Bureau of Economic Research, 2006.

Allmendinger, J. Educational Systems and Labour Market Outcomes. *European Sociological Review*, 5:231–250, 1989.

Amable, B., Demmou, L., and Gatti, D. Employment Performance and Institutions: New Answers to an Old Question, Discussion Paper No 2731, IZA, Bonn, Germany, 2007.

Angrist, J. D. and Pischke, J. S. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press, 2009.

Autor, D. H., Donohue, J. J., and Schwab, S. J. The costs of wrongful discharge laws. *The Review of Economics and Statistics*, 88:211–231, 2006.

Avdagic, S. Does deregulation work? Reassessing the unemployment effects of employment protection. *British Journal of Industrial Relations*, 53(1): 6–26, 2015.

Baccaro, L. and Rei, D. Institutional determinants of unemployment in OECD countries: does the de-regulatory view hold water? . *International Organization*, 61(3):527–569, 2007.

Banerji, A., Lin, H., and Saksonovs, M. Youth Unemployment in Advanced Europe: Okun's Law and Beyond. IMF Working Paper, Washington, DC, 2015.

Barbieri, P. Flexible employment and inequality in Europe. *European Sociological Review*, 25:621–628, 2009.

Beck, N. and Katz, J. Modelling dynamics in time-series-cross-section political economy data. *Annual Review of Political Science*, 14:331–52, 2011.

Belot, M. and van Ours, J. C. Does the Recent Success of Some OECD Countries in Lowering their Unemployment Rates Lie in the Clever Design of their Labor Market Reforms? . *Oxford Economic Papers*, 56(4):621–642, 2004.

Bentolila, S. and Bertola, G. Firing costs and labour demand: how bad is eurosclerosis? *The Review of Economic Studies*, 57(3):381–402, 1990.

Bertola, G., Blau, F. D., and Kahn, L. M. Labour market institutions and demographic employment patterns. *Journal of Population Economics*, 20 (4):833–867, 2007.

Blanchard, O. and Landier, A. The perverse effects of partial labour market reform: fixed-term contracts in France. *The Economic Journal*, 112:214–244, 2002.

Blanchard, O. and Wolfers, J. The role of shocks and institutions in the rise of European unemployment: the aggregate evidence. *The Economic Journal*, 110(462):C1–C33, 2000.

Blanchflower, D. G. and Freeman, R. B. *The Declining Economic Status of Young Workers in OECD Countries. In Blanchflower, D.G. & Freeman, R. B. (Eds)*. Youth Employment and Joblessness Advanced Countries, IL: University of Chicago Press, pp. 19-56 , 2000.

Blau, F. D. and Kahn, L. M. *Institutions and laws in the labor market. In Ashenfelter, O.C. & Card, C. (Eds.), Handbook of Labor Economics*. Amsterdam, Elsevier, pp. 1399–1461, 1999.

Blossfeld, H. P., Hofäcker, D., Hofmeister, H., and Kurz, K. Globalisierung, Flexibilisierung und der Wandel von Lebensläufen in modernen Gesellschaften. In Flexibilisierung (pp. 23-46). VS Verlag für Sozialwissenschaften, 2008.

Boeri, T. and van Ours, J. C. *The Economics of Imperfect Labor Markets*. Princeton and Oxford, Princeton University Press, 2008.

Botero, J. C., Djankov, S., Porta, R. L., Lopez-de Sianes, F., and Shleifer, A. The regulation of labor. *The Quarterly Journal of Economics*, 119(4): 1339–1382, 2004.

Breen, R. Explaining Cross-National Variation in Youth Unemployment: Market and Institutional Factors. *European Sociological Review*, 21(2):125–34, 2005.

Breen, R. and Buchmann, M. Institutional Variation and the Position of Young People: A Comparative Perspective. *Annals of the American Academy of Political and Social Science*, 580(1):288–305, 2002.

Brzinsky-Fay, C. The interplay of educational and labour market institutions and links to relative youth unemployment. *Journal of European Social Policy*, 27(4):346–359, 2017.

Bukodi, E. and Robert, P. Occupational mobility in Europe. European Foundation for the Improvement of Living and Working Conditions, Luxembourg: Office for Official Publications of the European Communities, 2007.

Busemeyer, M. R. and Iversen, T. Collective skill systems, wage bargaining, and labor market stratification. *The political economy of collective skill formation*, pages 205–233, 2012.

Cahuc, P. and Postel-Vinay, F. Temporary jobs, employment protection and labour market performance. *Labor Economics*, 9(1):63–91, 2002.

Cahuc, P., Carcillo, S., and Zylberberg, A. *Labor Economics*. Cambridge and London: MIT Press, 2004.

Cahuc, P., S., C., Rinne, U., and Zimmermann, K. F. Youth Unemployment in Old Europe: The Polar Cases of France and Germany. *IZA Journal of European Labor Studies*, 2(1):18, 2013.

Caroleo, F. E., Ciociano, E., and Destefanis, S. Youth Labour-Market Performance, Institutions and Vet Systems: A Cross-Country Analysis. *Italian Economic Journal*, 3(1):39–69, 2017.

Cazes, S. and Nesporova, A. *Labour Markets in Central and South-Eastern Europe: From Transition to Stabilization. In Cazes, S. & Nesporova, A. (eds). Flexicurity: A Relevant Approach in Central and Eastern Europe*. Switzerland, ILO, 2007.

Cazes, S. and Tonin, M. Employment protection legislation and job stability: a European cross-country analysis. *International Labour Review*, 149(3): 261–285, 2010.

Centeno, M. and Novo, A. A. Excess worker turnover and fixed-term contracts: causal evidence in a two-tier system. *Labor Economics*, 19(3): 320–328, 2012.

Della Porta, D. 11 Comparative analysis: case-oriented versus variable-oriented research. *Approaches and methodologies in the social sciences*, page 198, 2008.

Di Tella, R. and MacCulloch, R. The Consequences of Labour Market Flexibility: Panel Evidence Based on Survey Data. *European Economic Review*, 49(5):1225–1259, 2005.

Dieckhoff, M. and Steiber, N. Institutional reforms and age-graded labour market inequalities in Europe. *International Journal of Comparative Sociology*, 53(2):97–119, 2012.

DiPrete, T. A. Life Course Risks, Mobility Regimes, and Mobility Consequences: A Comparison of Sweden, Germany, and the United States. *American Journal of Sociology*, 108(2):267–309, 2002.

DiPrete, T. A., Goux, D., Maurin, E., and Quesnel-Vallee, A. Work and pay in flexible and regulated labor markets: A generalized perspective on institutional evolution and inequality trends in Europe and the US. *Research in Social Stratification and Mobility*, 24(3):311–332, 2006.

Eichhorst, W., Feil, M., and Braun, C. What Have We Learned? Assessing Labour Market Institutions and Indicators. IAB-Discussion Paper No 3470, 2008. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1294553. Viewed on June 27, 2019.

Eichhorst, W., Hinz, T., Marx, P., Peichl, A., Pestel, N., Siegloch, S., and Tobsch, V. Geringfügige Beschäftigung: Situation und Gestaltungsoptionen. Gütersloh, Bertelsmann Stiftung, 2012. URL http://ftp.iza.org/report_pdfs/iza_report_47.pdf. Viewed on July, 3, 2019.

Eichhorst, W., Hinte, H., and Rinne, U. Youth Unemployment in Europe: What to do about it? IZA Policy Paper No 65. Leibniz Information Centre for Economics, 2013.

Eichhorst, W., Rodríguez-Planas, N., Schmidl, R., and Zimmermann, K. F. A roadmap to vocational education and training systems around the world, 2015. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2199793. Viewed on July 3, 2019.

Esping-Andersen, G. *Who Is Harmed by Labour Market Regulations? Quantitative Evidence. In Esping-Andersen, G. & Regini, M- (eds). Why Deregulate Labour Markets?* . Oxford, Oxford University Press, 2000.

European Commission. Employment in Europe, 2006. Brussels: European Commission, 2006.

European Commission. Annual Growth Survey 2012. Brussels: European Commission, 2011.

Eurostat. Youth in Europe: A statistical portrait. Luxembourg: Publications Office of the European Union, 2009.

Feldmann, H. The Unemployment Effects of Labor Regulation around the World. *Journal of Comparative Economics*, 37(1):76–90, 2009.

Fialová, K. and Schneider, O. Labor Market Institutions and Their Effect on Labor Market Performance in the New EU Member Countries. *Eastern Europe an Economics*, 47(3):57–83, 2009.

Gangl, M. Changing labour markets and early career outcomes: labour mar-

ket entry in Europe over the past decade. *Work, Employment and Society*, 16(1):67–90, 2002.

Gangl, M. The Only Way Is up? Employment Protection and Job Mobility among Recent Entrants to European Labour Markets. *European Sociological Review*, 19(5):429–449, 2003.

Gash, V. and McGinnity, F. Fixed-term contracts – the new European inequality? Comparing men and women in West Germany and France. *Socio-Economic Review*, 5(3):467–496, 2007.

Gebel, M. Fixed-term contracts at labour market entry in West Germany: implications for job search and first job quality. *European Sociological Review*, 25(6):661–675, 2009.

Gebel, M. and Giesecke, J. Labor market flexibility and inequality: the changing skill-based temporary employment and unemployment risks in Europe. *Social Forces*, 90(1):17–40, 2011.

Gebel, M. and Giesecke, J. Does Deregulation Help? The Impact of Employment Protection Reforms on Youths' Unemployment and Temporary Employment Risks in Europe. *European Sociological Review*, 32(4):486–500, 2016.

Giesecke, J. and Groß, M. Temporary employment: chance or risk? *European Sociological Review*, 19(2):161–177, 2003.

Griffith, R., Harrison, R., and Macartney, G. Product Market Reforms, Labour Market Institutions and Unemployment. *Economic Journal*, 117(519): C142–C166, 2007.

Hadjivassiliou, K., Kirchner Sala, L., and Speckesser, S. Key indicators and drivers of youth unemployment. Institute for Employment Studies (IES), WP3-Policy Performance and Evaluation Methodologies, Version-2.0, 2015. URL https://www.style-research.eu/wp-content/uploads/ftp/STYLE-Working-Paper-WP3_1.pdf. Viewed on July 6, 2019.

Hannan, D., Raffe, D., and Smyth, E. Cross-National Research on School to Work Transitions: An Analytical Framework. Background paper prepared for the Planning Meeting for the Thematic Review of the Transition from Initial Education to Working Life. La Ciotat, France, 1996.

Hanushek, E. A., Woessmann, L., and Zhang, L. General Education, Vocational Education, and LaborMarket Outcomes over the Lifecycle. NBER Working Paper 17504, 2011.

Honaker, J. and King, G. What to Do about Missing Values in Time-Series Cross-Section Data. *American Journal of Political Science*, 54(2):561–581, 2010.

Honaker, J., King, G., and Blackwell, M. AMELIA II: A Program for Missing Data. *Journal of Statistical Software*, 45(7):1–54, 2011.

Jimeno, J. J. and Rodriguez-Palenzuela, R. P. Youth Unemployment in the OECD: Demographic Shifts, Labour Market Insti-tutions, and Macroeconomic Shocks. European Central Bank Working Paper No. 155. Frankfurt: European Central Bank, 2002.

Julkunen, I. *Youth Unemployment and Marginalization. In Furlong, A. (ed.), Handbook of Youth and Young Adulthood: New Perspectives and Agendas*. London, New York, Routledge, 2010.

Kahn, L. M. Employment protection reforms, employment and the incidence of temporary jobs in Europe: 1996 – 2001. *Labor Economics*, 17(1):1–15, 2010.

Kahn, L. M. Labor market policy: a comparative view on the costs and benefits of labor market flexibility. *Journal of Policy Analysis and Management*, 31(1):94–110, 2012.

Kerckhoff, A. C. Institutional Arrangements and Stratification Processes in Industrial Societies. *Annual Review of Sociology*, 21(1):323–347, 1995.

Kerckhoff, A. C. *Transition from School to Work in Comparative Perspective. In Hallinan, M.T. (ed.). Handbook of the Sociology of Education*. New York, Kluwer Academic, 2000.

King, G., Keohande, R. O., and Verba, S. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. NJ: Princeton, Princeton University Press, 1994.

Korpi, T. and Levin, H. Precarious footing. Temporary employment as a stepping stone out of unemployment in Sweden. *Work, Employment, Society*, 15(1):127–148, 2001.

Kugler, A. D. and Pica, G. Effects of employment protection on worker and job flows: evidence from the 1990 Italian reform. *Labor Economics*, 15 (1):78–95, 2008.

Levels, M., van der Velden, R., and di Stasio, V. From School to Fitting Work: How Educationto- Job Matching of European School Leavers Is Related to Educational System Characteristics. *Acta Sociologica*, 57(4):341–61, 2014.

Lijphart, A. Comparative Politics and the Comparative Method. *American Political Science Review*, 65(3):682–93, 1971.

Marinescu, I. Job security legislation and job duration: evidence from the United Kingdom. *Journal of Labor Economics*, 27(3):465–486, 2009.

Maurice, M., Sellier, F., and Silvestre, J. J. *The Social Foundations of Industrial Power: A Comparison of France and Germany*. Cambridge, MA: MIT Press, 1986.

McGinnity, F., Mertens, A., and Gundert, S. A bad start? Fixed-term contracts and the transition from education to work in West Germany. *European Sociological Review*, 21(4):359–374, 2005.

Müller, W. and Shavit, Y. *The Institutional Embeddedness of the Stratification Process: A Comparative Study of Qualifications and Occupations in Thirteen Countries. In Shavit, Y. & Müller, W. (Eds). From School to Work: A Comparative Study of Educational Qualifications and Occupational Destinations*. Okford, UK: Clarendon Press, 1998.

Nickell, S. Unemployment and Labor Market Rigidities: Europe versus North America. *Journal of Economic Perspectives*, 11(3):55–74, 1997.

Nickell, S. and Layard, R. *Labor market institutions and economic performance. In Ashenfelter, O. C. & Card, C. (Eds.), Handbook of Labor Economics*. Amsterdam, Elsevier, 1999.

Noelke, C. Employment protection legislation and the youth labour market. *European Sociological Review*, 32(4):471–485, 2015.

O'higgins, N. Youth unemployment and employment policy: A global perspective, 2001. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3019316. Viewed on July 8, 2019.

Organisation for Economic Cooperation and Development. Employment Outlook. Paris: OECD, 1999.

Organisation for Economic Cooperation and Development. Employment Outlook. Paris: OECD, 2004.

Organisation for Economic Cooperation and Development. Employment Outlook. Paris: OECD, 2006.

Organisation for Economic Cooperation and Development. Education at a Glance. Paris: OECD, 2013.

Organisation for Economic Cooperation and Development. Education at a Glance. Paris: OECD, 2014.

Organisation for Economic Cooperation and Development. Education at a Glance. Paris: OECD, 2019.

Pohl, A. and Walther, A. Activating the disadvantaged. Variations in addressing youth transitions across Europe. *International Journal of Lifelong Education*, 26(5):533–553, 2007.

Polavieja, J. G. Temporary contracts and labour market segmentation in Spain. An employment-rent approach. *European Sociological Review*, 19 (5):501–517, 2003.

Quintini, G. and Manfredi, T. Going separate ways? School-to-work transitions in the United States and Europe. *OECD Social, Employment and Migration Working Papers*, 90:1–53, 2009.

Quintini, G., Martin, J. P., and Martin, S. The Changing Nature of the School-to-Work Transition Process in OECD Countries. *IZA Discussion Paper (No. 2582)*, 2007.

Rovelli, R. and Bruno, R. Labor Market Policies, Institutions and Employment Rates in the EU-27. Discussion Paper No 3502. Bonn, Germany: IZA, 2008.

Russell, H. and O'Connell, P. Getting a Job in Europe: The Transition from Unemployment to Work among Young People in Nine European Countries. *Work, Employment & Society*, 15(1):1–24, 2001.

S., B. and Dolado, J. J. Labour flexibility and wages: lessons from Spain. *Economic Policy*, 9(18):53–99, 1994.

Salvador, R. G. and Leiner-Killinger, N. An analysis of youth unemployment in the euro area. *ECB Occasional Paper: European Central Bank*, 2008.

Scherer, S. Patterns of labour market entry: long wait or career instability? An empirical comparison of Italy, Great Britain and West Germany. *European Sociological Review*, 21(5):427–440, 2005.

Schönmann, I. and Clauwaert, S. The Crisis and National Labour Law Reforms: A Mapping Exercise. ETUI Working Paper 04. Brussels: European Trade Union Institute, 2012.

Spence, M. Job Market Signalling. *The Quarterly Journal of Economics*, 87: 355–374, 1973.

StataCorp. Stata: Release 14, 2015.

Stigler, G. J. Information in the Labor Market. *Journal of Political Economy*, 70(5):94–105, 1962.

Stiglitz, J. E. A Theory of "Screening" Education, and the Distribution of Income. *American Economic Review*, 65(3):283–300, 1975.

Van Der Velden, R. and Wolbers, M. H. J. *The Integration of Young People into the Labour Market: The Role of Training Systems and Labour Market Regulation. In Müller, W. & Gangl M. (eds). Transitions from Education to Work in Europe: The Integration of Youth into EU Labour Markets*. Oxford and New York, Oxford University Press, pp. 174–99, 2003.

Venn, D. Legislation, Collective Bargaining and Enforcement: Updating the OECD Em-ployment Protection Indicators. OECD Social, Employment and Migration Working Paper No. 89. Paris: OECD, 2009.

Wolbers, M. H. J. Patterns of Labour Market Entry: A Comparative Perspective on School-to-Work Transitions in 11 European Countries. *Acta Sociologica*, 50(3):189–210, 2007.

World Bank. World Development Report 2007: Development and the Next Generation. *DC: The World Bank*, 2007.

Zimmermann, K. F., Biavaschi, C., Eichhorst, W., Giulietti, C., Kendzia, M. J., Muravyev, A., Peters, J., Rodríguez-Planas, N., and Schmidl, R. Youth Unemployment and Vocational Training. *Foundations and Trends in Microeconomics*, 9(1):1–157, 2013.

Junior Management Science 6(1) (2021) 81-99

Junior Management Science

journal homepage: www.jums.academy

The Impact of Management, Family and Employee Ownership Concentration on Firm Performance

Michael Amroudi

Technische Universität München

Abstract

This thesis investigates the relation between ownership structure and firm performance using a sample of 2,120 publicly traded European companies. The question of whether this relation should be positive or negative has been the subject of a wide-ranging discussion and was addressed by many researchers. Of particular interest have been management, family, and employee owners. Nevertheless, there is no consensus in the literature, and empirical studies on European companies are scarce.

Utilising data from the European Federation of Employee Share Ownership (EFES) and the Bureau Van Dijk Orbis database, this relation is analysed using multiple linear regression with continuous and categorical predictors. The results show that firms having a management owner concentration up to strategic levels report a significantly higher Tobin's Q than firms having a no or no significant management concentration. The same effect holds true for family owners and employee owners. Measuring the ownership structure as the fraction of shares of the largest shareholder does not yield a significant effect and highlights the value of management, family, and employee owners.

Keywords: Ownership concentration; family ownership; management ownership; employee ownership; firm performance.

1. Introduction

Ownership structures and their implications have been a widely discussed topic for almost a century and are still in the interest of the finance and management literature. Beginning with the argument by Berle and Means[1] in 1932 that an increasing separation of ownership and control will lead to a long-term deterioration in the value of a company, numerous other arguments for and against it have developed.

At the same time, ownership structures are changing worldwide. Employee Share Ownership Plans (ESOP), for example, are gaining popularity in the United States. But in Europe, too, growing numbers of financial participation schemes can be observed. Between 1999 and 2005, the share of companies offering broad-based ownership schemes rose from 13% to 18%.[2] In the case of profit-sharing plans, the share rose from 29% to 35%.[3]

However, the study of ownership structures is also of macroeconomic relevance. Family owners, in particular, have now accumulated an intangible value that can have a significant impact on the economy as a whole. In Sweden, for example, a single family, the Wallenbergs, controls about half of the market capitalization of the Stockholm Stock Exchange.[4] In Hong Kong, the most valuable 15 families control 84.2% of the gross domestic product.[5]

But is a higher owner concentration good for the performance of a company? There is no empirical consensus on that in the literature, and most studies focus on an isolated analysis of owner types. Most often discussed are management, family, and employee owners. The results, however, remain divided: Management owners should have positive effects in high concentrations because they combine control and ownership, thereby reducing agency costs.[6] However, more management owners also lead to a higher risk of entrenchment, allowing large shareholders to make decisions at the expense of minority shareholders.[7] Families should have a good impact on companies because they bring finan-

[1] See Berle and Means (1932).
[2] See Lowitzsch, Hashi, Woodward, and (Eds.) (2009), p. 22.
[3] See Lowitzsch et al. (2009), p. 23.
[4] See Agnblad, Berglöf, Högfeldt, and Svancar (2001).
[5] See Claessens, Djankov, and Lang (2000), pp. 107-109.
[6] See for example Ang, Cole, and Lin (2000), p. 83-84.
[7] See for example Villalonga and Amit (2006), p. 387.

DOI: https://doi.org/10.5282/jums/v6i1pp81-99

cial access and have an above-average incentive to monitor the company.[8] On the other hand, they are often accused of abusing their power and limiting management positions to family members.[9] Finally, Employee Owners should also have a positive influence on the company, as financial participation increases their motivation.[10] But the opposite may also be true, as more labour participation also leads to inefficient decision-making processes.[11]

This thesis analyses the relationship between three important owner types, i.e. management, family and employee owners, and firm performance. It is argued that a higher concentration of each of these owner types should lead to better firm performance. The results of several multiple linear regressions show a significant positive correlation up to a concentration level where strategic ownership is exceeded. An increase in owner concentration, therefore, is most beneficial up to a share of 20%.

Most existing studies are using exclusively American companies in their sample. This study focuses on European companies and thus contributes to the current literature by applying similar models to a less investigated environment. At the same time, this work breaks the frequently made assumption of a linear relationship and provides indications that there is an optimal concentration level for each owner. Practitioners conclude from this strategic implication that suboptimally structured firms can improve the performance of their firms by increasing the concentration of certain owner groups. What this study will not do is to compare different types of owners.

The rest of this work is structured as follows: Chapter 2 presents a set of criteria for characterizing ownership structures. By comparing the most common arguments, hypotheses about the influence owner concentrations should have on the company performance are derived. Chapter 3 introduces the methodology used in this study and describes the variables used and the regression models. The results of the regressions are then presented in Chapter 4 and evaluated in relation to the initial question. Since the models may have some weaknesses, possible issues and striking results are discussed further in chapter 5, where the robustness of the results is tested with the help of further models.

2. Ownership Structure and Firm Performance

2.1. Classifying Ownership Structures

In order to investigate the influence of ownership structures on firm performance, a clear method for differentiating between different structures must first be defined. The ownership structure of a company can be described under three different aspects: The identity of the owner, his share in the company, and finally, his involvement in the company.

The identity of an owner refers to the type of a shareholder and includes managers, families, employees, governments, banks, foreign investors and other companies. A classification by ownership identity is relevant as different owners can differ in terms of their interests, risk preferences, time horizons and strategies.[12] Shareholder value is not a universal goal of corporate strategy. For example, while family ownership may be associated with the goal of firm survival, state ownership is expected to pursue political goals such as low output prices and low unemployment rates.[13] Moreover, this also implies that hybrid ownership structures with multiple identities can be related to diverging interests.

The second aspect, ownership concentration, is determined by the proportion of an owner's shares in a company and is often divided into concentrated ownership and dispersed ownership.[14] It is not clear which threshold should be used to differentiate between these two types of ownership. There is agreement in the literature, however, that a stake of less than 50 per cent may be sufficient to give a block holder ultimate control over the company.[15] Often a threshold of 20 or 25 per cent is applied as a definition for concentrated ownership.[16] Accordingly, a structure in which no shareholder owns more than 20 per cent of the company's shares is defined as dispersed ownership.

A third aspect to describe the ownership structure of a company is the involvement of the owner which can be measured, e.g. as the number of an owner's votes or as the number of top executive positions filled by members of a certain identity. A typology to classify the interaction between the owner and the company is to differentiate between relational and transactional owners.[17] A relational owner is a long-term oriented shareholder with complex profit and growth goals for a company. Such a characteristic can often be found in families. A transactional owner, on the other hand, does not engage in further relations with the company except for obtaining returns. Such a characteristic can be often found in institutional investors.

As ownership involvement is difficult to measure, this aspect is neglected in most studies, and so it will be in this thesis. Hence, the ownership identity and concentration are the two most widely used aspects to classify a company's ownership structure.

2.2. Literature Review

The relationship between different ownership structures and firm performance has been the subject of numerous studies since Berle and Means[18] and has led to a widely discussed

[8] See Anderson and Reeb (2003), pp.1306-1307.
[9] See Anderson and Reeb (2003), pp.1301-1302.
[10] See for example Mazibuko and Boshoff (2003), p. 31, Kurtulus and Kruse (2017), p. 1.
[11] See Park, Kruse, and Sesil (2004), pp. 5-7.
[12] See Aguilera and Jackson (2003), pp. 450f.
[13] See for example Lins, Volpin, and Wagner (2013); Thomsen and Pedersen (2000).
[14] See Pedersen and Thomsen (1997), pp. 765-769.
[15] See Faleye, Mehrotra, and Morck (2006), p. 7.
[16] See for example Aguilera and Crespi-Cladera (2016); Lins et al. (2013); Pedersen and Thomsen (1997).
[17] See for example Aguilera and Jackson (2003); David, O'Brien, Yoshikawa, and Delios (2010), p. 638.
[18] See Berle and Means (1932).

disagreement about whether and what kind of relationship exists. Most studies have focused on an isolated consideration of individual owner identities and arguments for and against each individual identity emerge. The discourse on three of the most important ownership structures, i.e. management ownership, family ownership and employee ownership, will be the subject of this section. Based on the arguments presented, hypotheses about the relationship between ownership structures and firm performance will be derived.

2.2.1. Management Ownership and Firm Performance

An important and frequently cited argument that predicts a positive impact of managerial ownership on firm performance is the agency theory presented by Jensen and Meckling[19]. It analyses the relationship between a principal and an agent, with the principals representing the owners of the firm and the agents representing the managers. The assumption is that owners pursue the goal of maximizing the company's profit and thus, shareholder value. Opposed to this are the interests of managers who strive for personal perks, such as high compensation. Agency costs arise when the interests of those who run a company and have decision-making authority do not coincide with the interests of the company owners. Accordingly, companies that are wholly owned by the management have, by definition, no agency costs and thus an optimal match between principal and agent. At the other extreme are managers who do not own any shares in the company. Higher management participation should, therefore, lead to higher company success. The notion that agency costs vary inversely with the manager's ownership share was examined and confirmed by Ang, Cole and Lin[20] who took – as the theory requires – non-publicly traded firms into account in their study.

This theory is consistent with numerous empirical studies that show a positive correlation between a higher concentration of management owners and corporate success. Oswald and Jahera[21], for example, found a significant positive influence of officer's and director's stockholdings on performance variables such as excess stock return, return on assets (ROA) and return on equity (ROE) in a sample of 645 companies listed on the New York and American Stock Exchange.

Core and Larcker[22] argue that there is an equilibrium where managers are optimally involved in the company. They have focused their sample on companies that have introduced a target ownership plan, which stipulates that managers own a certain minimum number of shares in the company. They observed statistically higher stock price returns in the first six months of a fiscal year in which such a plan was announced. In addition, these companies showed statistically higher excess accounting returns in the two years following the introduction of the plan. An increase in the management's stake in the company, who have a below-equilibrium percentage of shares, therefore has a positive impact on performance.

On the other hand, the theory that a higher proportion of management owners leads to a higher company valuation has been challenged by the entrenchment hypothesis.[23] Managers who own at least a certain percentage of the company can become entrenched and exploit this situation at the expense of smaller shareholders.[24] More management ownership should, therefore, also lead to more entrenchment. A study by Morck, Schleifer and Vishny[25], however, does not find a clear answer to this hypothesis. Although Tobin's Q decreases at a management concentration between 5 and 25 per cent compared to the concentration category between 0 and 5 per cent, as soon as the management ownership exceeds 25 per cent, Tobin's Q increases again, which argues against entrenchment.

Demsetz[26] presented another argument that questions the relationship between management ownership and firm performance. The agency theory – also called the convergence-of-interest hypothesis – is based on the false assumption that all managers primarily strive for higher corporate profitability. However, since not all managers have the same objectives, they will work only for companies that fit their preferences. At the same time, companies compensate their managers to adjust for the agency costs incurred which, eventually, will lead to an optimal balance between management and non-management owners in every company. The ownership structure of a company is thus an endogenous consequence of the decisions of current and future shareholders and should therefore not have any influence on the company. For example, in a non-optimal dispersed ownership structure, a company would, if necessary, carry out an appropriate concentration of management shares itself to compensate for its ineffectiveness.

Empirical evidence of this idea can be found in studies by Demsetz and Villalong and, Yixiang.[27] Both conclude that there is no link between management shares and business success when the ownership structure is modelled as an endogenous variable. But even this approach is not consistently found in the literature. Kapopoulos and Lazaretou[28], for example, have applied the same model and for their sample have shown that a higher concentration of management owners is associated with higher firm performance. They looked at a sample of 175 companies listed in Greece and measured performance using ROA and Tobin's Q.

Based on the argumentation presented and the evidence hinted at in the literature, it can be concluded that there is still no clear position on the relationship between these two variables. However, since a tendency towards a positive influence is emerging, the following hypothesis can be made:

[19] See Jensen and Meckling (1976), pp. 312-319.
[20] See Ang et al. (2000), pp. 81-83.
[21] See Oswald and Jahera Jr (1991), pp 324f.
[22] See Core and Larcker (2002), p. 321, p. 338.
[23] See Morck, Shleifer, and Vishny (1988), pp. 294f.
[24] See Thomsen and Pedersen (2000), p. 690.
[25] See Morck et al. (1988), pp.299-302.
[26] See Demsetz (1983), pp. 382-385.
[27] See Demsetz and Villalonga (2001); Yixiang (2011).
[28] See Kapopoulos and Lazaretou (2007), pp. 149-153.

Hypothesis 1: Higher ownership concentration of executive employees (managers) is positively related to firm performance.

2.2.2. Family Ownership and Firm Performance

The second owner identity frequently discussed in the literature are families. In terms of the agency theory mentioned above, family businesses should do well, since owners and managers are often either the same persons or at least have a close relationship with each other.

In contrast to this, however, there is the view that families differ from typical managers in at least one essential characteristic: Unlike managers, who primarily aim for the profitability of the company, families have an interest in keeping the company going for many generations. This leads to a bias towards survival-oriented actions so that families are considered to make decisions that can come at the expense of other shareholders.[29] An example is the tendency of family businesses to limit executive management positions exclusively to family members. This limited pool of potential, qualified candidates can lead to a competitive disadvantage compared to non-family businesses.[30] Another example can be found in the investment decisions of family owners. In the course of an economic recession, family businesses reduce their investments much more than other companies, which leads to a stronger decline in stock prices.[31] According to Lins, Volpin and Wagner[32], family firms tend to underperform by 2.0 to 3.3 percentage points in a financial crisis. Moreover, when a family holds multiple companies and one of them is hit by a crisis, it can spread across the other healthy firms. It is, however, questionable to what extent these results can be generalized to other periods of the economic cycle.

This risk of power abuse by family managers implies that companies that are owned but not managed by a family should perform better than companies in which the family is actively involved in the management. Hence, the agency theory should not hold. Barotini and Caprio[33], however, provide evidence that when families take executive positions, they cannot be distinguished from non-family firms and perform at least as good as their peer groups. However, when they are not represented on the board, family firms perform worse than non-family firms. This result invalidates the argument that families should not be involved in the management of a firm.

In addition, an interest in the long survival of the company can also be interpreted just as well to the advantage of family businesses. The historical presence of the families, as well as their often non-diversified company portfolio and their control of management positions, provide an exceptionally good basis for monitoring of the company. The incentives of family owners to monitor more closely arise from the fact that the consequences of their actions directly return to them.[34] In addition, families face reputational damage in the event of poor performance.[35] There are also further advantages of the long-term orientation of families: James[36] argues that decisions made by managers who are members of the founding family are generally more efficient than in other companies. This is because their decisions are based on an altruistic foundation that takes into account the consequences for the entire family and also for future family members. For example, owner-managers of a family make more efficient investment decisions than comparable non-family managers when they know that they will one day hand over the business to another family member.

Another advantage of family businesses is their financial access. According to Morck, Wolfenzon and Yeung[37], most family businesses exist in a pyramid structure in which a family controls several companies, which in turn, control other companies. Companies in a pyramid control structure report higher ROA. In addition, they are often considered to have easier access to financing, as they can benefit from a common pool of resources which is created as a result of other firms under the family's control. Furthermore, family firms have a significantly lower cost of debt, which implies that bondholders trust family firms to protect their interests better, e.g. because of long-term survival.[38]

A study by Andersond and Reeb[39] concludes that family firms perform significantly better than non-family firms in terms of Tobin's Q and ROA. The result of the 403 firm sample is not only statistically but also economically significant since Tobin's Q of family firms is ten per cent higher. Numerous other studies show a positive influence of family owners on firm performance.[40] Thomsen and Pedersen[41], on the other hand, prove a negative influence on a data set of 435 public companies. According to Martínez, Stöhr and Quiroga[42], the negative influence identified by some studies is due to the fact that these studies only looked at public companies which generally perform worse. Relevant to the study of family ownership, on the other hand, are private companies that perform significantly better.

In summary, stronger arguments and more empirical evidence for a positive influence of family owners on the success of a company can be found. The following hypothesis emerges from the literature review presented here:

Hypothesis 2: Higher ownership concentration of founding family members is positively related to firm performance.

[29] See Lins et al. (2013), p. 2584; Anderson and Reeb (2003), pp. 1306f.
[30] See Anderson and Reeb (2003), pp. 1301f; Barontini and Caprio (2006), p. 690.
[31] See Lins et al. (2013), pp. 2595-2598.
[32] See Lins et al. (2013), pp. 2595-2598.
[33] See Barontini and Caprio (2006), pp. 720f.
[34] See Silva and Majluf (2008), p. 610.
[35] See Anderson and Reeb (2003), p. 1306.
[36] See James (1999), p. 44, p. 52.
[37] See Morck, Wolfenzon, and Yeung (2005), p. 661.
[38] See Anderson and Reeb (2003), p. 1306.
[39] See Anderson and Reeb (2003), pp. 1317-1322.
[40] See for example Martínez, Stöhr, and Quiroga (2007); Villalonga and Amit (2006).
[41] See Thomsen and Pedersen (2000).
[42] See Martínez et al. (2007), pp. 84-86.

2.2.3. Employee Ownership and Firm Performance

The third important owner identity to be investigated are employees. In both the US and Europe, employee financial participation has become increasingly important and popular. As stated in the report on the "Promotion of Employee Participation in Profits and Enterprise Results" (known as the PEPPER recommendations) by the European Commission, employee participation is tax-incentivized. This incentive is crucial in explaining the increase in the number of Employee Share Ownership Plans (ESOP), a form of employee ownership.[43] However, there are several other arguments that try to explain the interaction between employee ownership and firm performance.

It comes naturally to mind that employee share ownership should have an impact on their motivation. However, it is less clear whether this should be a positive or negative impact. One way of looking at this question is to model this as a game-theoretical problem.[44] The idea behind this argument is that the participation of employees in the company's success does not generate an incentive to work more since employees can profit from the success of others even without extra work. According to the logic of the "Prisoner's Dilemma", employees can generate more profit in absolute terms by working together cooperatively. However, they can generate more personal utility if they shirk and rest on the work of others. So, employees have an incentive to be a free rider and this incentive grows with the size of the group. This increasing incentive to shirk or the decreasing incentive to cooperate is also often called the 1/N problem because a worker in a group with N members receives only 1/N of the additional reward earned.[45] This becomes problematic if all employees follow this view because then all employees will get to the worst situation with minimal profit.

However, following Pendleton[46], this game-theoretical dilemma can be easily solved. On the one hand, there is a consensus in the literature that repeated playing motivates individuals to cooperate and pursue an optimal strategy. On the other hand, shirking from one employee leads to higher costs for all other employees. These costs would lead to a situation where employees who resist the expectation of working harder would be sanctioned by peer pressure, for example. An employee ownership structure, therefore, reduces the monitoring costs of a company. This view was confirmed in a study by Freeman, Kruse and Blasi[47], who evaluated the responses of over 41,000 employees from 14 companies. An anti-shirking index was calculated from the answers to the question of how one would react if one saw an employee working less than he or she should. The result shows that employees who own company shares are significantly more likely to respond to this observation actively.

The other view on the question of whether employee share ownership should have a positive or negative impact is that it will positively motivate employees. In fact, it has often been proven that such shares motivate employees to work harder to increase the value of their shares. In addition, they complain less, take advantage of more training opportunities, are absent less often, and there is generally lower labour turnover in the company.[48] Employee participation, therefore, leads to greater involvement of employees in the performance of the company they work for and in the implementation of their interests. Similar to family owners, employees have survival-oriented interests and additionally strive for employment stability.[49] It follows that companies with an employee ownership structure should survive longer than comparable companies with a different ownership structure. The empirical investigation of this hypothesis by Park, Kruse and Sesil[50] showed that companies that are at least five per cent owned by employees are only 76 per cent as likely to disappear in the same period as companies without employee ownership. This interest also leads to making certain decisions of the company more comprehensible. It is argued that employees are more willing to adapt and to accept change if the company is forced to do so by the economic situation.

The comparison of these two views, therefore, suggests a positive influence of employee share ownership on work motivation and is thus a first indication of a better performance of the company.

At this point, it should be noted that a higher interest of employees to become involved in the company can have a negative impact on other stakeholders. As will be shown later (see section 5.5.2), the interaction of different owners probably also plays a role in the evaluation of firm performance. This gives rise to the danger that too much employee involvement can have negative consequences for managers, for instance. In the literature, this is addressed by the influence of employee share ownership on decision making efficiency. Employee share ownership encourages employees to become too involved in the decision-making process, which leads to a collective decision-making problem and undermines managerial authority.[51] Yet, a higher level of involvement can just as well be understood positively. For example, it is not taken into account that it can also be helpful to include different competencies in the decision-making process. In fact, it is beneficial to involve employees in the company, as this encourages all levels of the hierarchy to consider how to improve the company. Frontline workers, who are often not involved in product process improvement plans, are experts in this field and their suggestions should be taken into account. In the long term, this will lead to an improvement in product quality and the production process.[52] It is also conceivable

[43] See Lowitzsch et al. (2009), pp. 13-21.
[44] See Park et al. (2004), pp. 5f.
[45] See for example Blasi, Freeman, and Kruse (2016); E. H. Kim and Ouimet (2014).
[46] See Pendleton (2002), pp. 108f.
[47] See Freeman, Kruse, and Blasi (2008), pp. 6-10.
[48] See for example E. H. Kim and Ouimet (2014); Kurtulus and Kruse (2017); Mazibuko and Boshoff (2003).
[49] See Kurtulus and Kruse (2017), p. 2.
[50] See Park et al. (2004), pp. 28-30.
[51] See for example Dow and Putterman (2000); Long (1978); Mazibuko and Boshoff (2003); Richter and Schrader (2017).
[52] See Park et al. (2004), p. 7.

that this effect can be mitigated by different implementations of the employee ownership structure.

Independent of work motivation and involvement, it is also possible to find arguments in favour of the general situation of the employees. Arguments that speak against the improvement of the employment situation are vanishingly small in the literature. Long[53] points out that employees can get the feeling that more work is expected of them through company shares. In the long term, this could damage productivity and the working climate. However, the sample used casts doubt on the generalizability of this result. Another point is that employees tend to have a poorly diversified portfolio due to the high proportion of shares in their own company.[54] Empirical evidence for this argument is not known.

In contrast, numerous studies prove a positive effect on employee attitudes. Kaarsemaker, Pendleton and Poutsma[55] reviewed more than 50 studies on this topic and summarised that more than two-thirds of them speak for a positive effect on attitudes. Frequent reasons for this result are that this builds up more trust in superiors and promotes pro-social behaviour as well as identification and responsibility with one's own company.[56] Companies with an Employee Ownership Structure consequently show a healthier corporate culture.

An empirical study such as the one that will be carried out here has already been dealt with frequently in the literature, albeit mainly in relation to American companies. Many studies conclude that a stronger concentration of employee owners is positively related to the financial performance of a company.[57] Fewer studies find empirical evidence to the contrary.[58] Based on these results and the discussion outlined above, it is expected that a similar effect can be found for European companies. Consequently, the next hypothesis follows:

> Hypothesis 3: Higher ownership concentration of non-executive employees is positively related to firm performance.

2.2.4. Concentrated Ownership and Firm Performance

In addition to managers, families and employees, companies can have numerous other shareholders who differ from one another in their initial motivation. Thomsen and Pedersen[59] make a broader analysis of various owners and compare their interests. Financial and insurance companies, for example, attach greater importance to shareholder value and portfolio diversification, corporate owners attach importance to business transactions and growth, and governments take social consequences into account in their actions in terms of jobs and social welfare. Even more extreme are foundation-owned companies, which are non-profit organizations without owner monitoring, and which pursue exclusively charitable purposes. Nevertheless, Thomsen and Rose[60] find that these companies are at least as efficient in terms of risk-adjusted stock returns, accounting returns and Tobin's Q compared to companies with other owner identities.

This raises the question of whether profit-seeking behaviour must be a necessary condition for a successful company. Corporate success could be a prerequisite for an owner to realize his individual interests so that each owner benefits from a higher firm performance. Consequently, the analysis of firm performance may be less concerned with the individual interests of the owner and more with his ability to realize these goals. A higher concentration of a single owner should result in that owner being able to realize his interests more efficiently than in an environment of many different shareholders with heterogeneous interests.

Theoretically, as discussed above, agency costs play an important role for block holders. In the literature, it is repeatedly emphasised that concentrated owners have economic incentives to minimize agency conflicts and actively monitor managers.[61]

Multiple studies provide evidence that a controlling block holder of any type has a significant positive correlation with firm performance.[62]

The arguments presented here, therefore, give rise to another hypothesis:

> Hypothesis 4: Firms having a concentrated ownership structure perform better than comparable firms with a dispersed ownership structure.

3. Methodology

In order to test the hypotheses empirically, this study conducts several multiple linear regressions using continuous and categorical predictors. The aim of this section is to describe the data used for the analysis as well as the structure of the applied models and their associated variables.

3.1. Data Sample

To investigate the effects of different owner identities on the performance of a company, a data set was compiled from three different databases.

The main data source is the database of employee share ownership in European companies from the European Federation of Employee Share Ownership (EFES).[63] It contains data from 31 European countries and comprises a panel of

[53] See Long (1978), p. 43.
[54] See Kurtulus and Kruse (2017), p. 3.
[55] See Kaarsemaker, Pendleton, and Poutsma (2009), p. 17.
[56] See for example Blasi et al. (2016); Richter and Schrader (2017).
[57] See for example Blasi et al. (2016); Haldar and Rao (2011); K. Y. Kim and Patel (2017); Park et al. (2004); Richter and Schrader (2017).
[58] See for example Faleye et al. (2006); Livingston and Henry (1980).
[59] See Thomsen and Pedersen (2000), p. 690.
[60] See Thomsen and Rose (2004), pp. 344f.
[61] See for example Demsetz (1983); Shleifer and Vishny (1986).
[62] See for example Lins et al. (2013); Thomsen and Pedersen (2000); Yixiang (2011).
[63] See EFES (2019).

2,709 companies from 2007 to 2016, including all listed companies with a stock market capitalisation of at least 200 million euros. It thus includes all significant European companies. The listed companies represent 25% of all European listed companies (excluding financial firms), but 99% in terms of capitalisation and 95% in terms of employment. In addition, non-listed enterprises with at least 100 employees and whose employees hold at least 50% of the company shares are included in the data set. The information in the EFES database is based on the annual reports provided by the companies themselves.

For the calculation of firm performance measures, the required financial data has been extracted from the Thomson Financial Worldscope database and attached to the companies contained in the EFES database.

Furthermore, the EFES dataset focuses exclusively on executive and non-executive employee ownership. From this, variables for determining employee and management ownership can be derived. To include additional information on family owners, the dataset was merged with the Bureau Van Dijk Orbis database, a continuously updated global database with extensive information on corporate ownership.

Finally, all companies without ISIN and all companies without employees were removed from the composite data sample. In addition, all financial firms (Groups 44 to 47) were removed using the Industry Portfolios by Fama and French[64]. The final sample contains 2,120 companies from 31 European countries. All numeric nonbinary variables have been winsorised at the 1st and 99th percentile to reduce the impact of outliers.

3.2. Variables

3.2.1. Firm Performance

In order to evaluate the performance of a company, researchers and practitioners utilize numerous different performance measures. This thesis focuses on Tobin's Q and ROA to evaluate the performance. The choice of these two variables is meaningful as they describe the company from two perspectives: market-based and accounting-based.

Tobin's Q is one of the most commonly used capital market performance measures.[65] It is calculated by dividing the market value of a company by the replacement cost of its assets based on the current market price of these assets. So, if the market value of the company is equal to its replacement cost, there is a balance between the market value and the intrinsic value, and Tobin's Q equals one. If the value is greater than one, the market attributes an additional value to the company. Tobin's Q can, therefore, also be interpreted as the expected future performance of a company. A higher Tobin's Q is consequently considered a better firm performance.

The second performance measure used in this analysis is the ROA ratio, which is calculated by dividing net income by total assets. It is an indicator of how profitable a company is relative to its assets. From this, it can be inferred how efficiently a company uses its available assets to generate earnings. Unlike Tobin's Q, ROA is an accounting-based performance measure and provides information about the historical performance of a company. ROA has also been widely used in the literature to examine the impact of ownership structures.[66]

3.2.2. Ownership Structure

As described in section 2.1, the ownership structure of a company is examined in this thesis based on ownership identity and owner concentration. In this analysis, it is not a question of comparing different owner types against each other in terms of their performance, but rather of examining the effect of the concentration of a single owner identity. For this reason, three separate analyses are carried out, one for management, one for families and one for employee owners.

When determining the owner identity, it should be noted that a distinction must be made between direct and indirect ownership. Family businesses, in particular, are often found to exist in a pyramidal control structure.[67] When using an indirect-ownership measure, a company that is owned by another company, which in turn in owned by a family, would be attributed to a family. Since indirect ownership is likely to have a different effect on company behaviour than direct ownership, only direct owners are considered in this study.

To determine the ownership concentration of an entity, multiple approaches are used. For management and employee owners, the concentration of the owner is measured based on their aggregated fraction of shares. The reason for this is that these owners are likely to pursue similar interests by taking comparable decisions (see sections 2.2.1 and 2.2.3). In families, on the other hand, the individual owners play a more important role. For this reason, their concentration is measured by the proportion of shares that the largest family member is holding. For the fourth hypothesis, the fraction of shares of the largest shareholder, who is commonly called 'Blockholder', is used.

The use of such continuous variables is based on the assumption of a linear relationship between the proportion of shares held by an owner and firm performance. However, some studies suggest that this relationship is non-linear. In order to adapt to these results, the ownership concentration is divided into five groups and represented by four dummy variables in an additional model. The first dummy variable covers ownership levels from at least 1% to less than 6%. The second runs from at least 6% to less than 20%, the third from 20% to 50% and the fourth variable comprises ownership levels of at least 50% of the total shares. According to Kim and Patel[68], the first group corresponds to significant ownership,

[64] See Fama and French (n.d.).

[65] See for example Anderson, Mansi, and Reeb (2003); Richter and Schrader (2017); Silva and Majluf (2008).

[66] See for example Barontini and Caprio (2006); Villalonga and Amit (2006); Xia and Walker (2015).

[67] See Morck et al. (2005).

[68] See K. Y. Kim and Patel (2017), p. 4.

the second to strategic ownership, the third to determining ownership and the fourth to controlling ownership.

Such a partition is too granular for blockholders, as the median blockholder owns 29.50% of all shares while, for example, the median management concentration in the sample is 0.18% (cp. Table 1). For this reason, two thresholds are used: One model uses a 25% threshold and the other one a 50% threshold.

3.2.3. Control Variables

Other potential determinants of firm performance are included as control variables. The first control variable is firm size which can be measured by the number of employees and by the market capitalization. As both of these variables are strongly correlated (Pearson correlation coefficient of 0.633), market capitalization is excluded from further analysis to reduce the impact of multicollinearity. Measuring firm size using the number of employees is found to fit this study design better because it is expected to be less volatile and less prone to other capital market interferences.

A firm's debt ratio can influence its net income which in turn affects the firm performance measures.[69] To account for this effect, leverage is included as a second control variable. It is calculated by dividing total debt by total assets.

To proxy for firm growth opportunities and future returns, research and development (R&D) intensity is also included as a control variable. It is calculated by dividing R&D expenditures by net sales. Including this variable is not common practice in the literature and has only been used by few researchers.[70] For this reason, R&D intensity is included in a separate model.

To control for year fixed effects, nine dummy variables are introduced representing the period range from 2008 to 2016. The dummies are 1 for a particular year and 0 otherwise. Furthermore, 347 dummy variables are added to control for an interaction of the country and the industry of a company.

3.3. Descriptive Statistics

Table 1 presents descriptive statistics for the dataset. By comparing the mean for the concentration variables, a strong positive skew can be noticed. As this also holds true for the performance measures and the control variables, all variables except for 'Blockholder Concentration', 'ROA' and 'R&D Intensity' are log-transformed in further analysis.

The median firm in this sample is rather small, with a market capitalisation of € 439 million and 2,545 employees. Half of the firms in this sample, have a Tobin's Q between 1.024 and 1.813 as well as ROA between 0.008 and 0.095, indicating interquartile ranges of 0.789 and 0.087, respectively.

It should be noted that there is a sharp decline in the number of observations for the control variable 'R&D Intensity'. The reason for this decline is a large number of missing values for R&D expenditures which may be caused by companies in which R&D does not take place. For example, companies from the textiles sector are less likely to undertake R&D investments than companies in the pharmaceutical industry.[71]

To further investigate the frequency of different owner concentration levels, Figure 1 provides a bar chart that depicts the number of firms in the sample that report corresponding ownership concentrations in 2016. It can be seen that the discrepancy in frequencies between different concentration levels is more pronounced for employee owners than it is for management or family owners. Of all sample firms in 2016, who report at least a significant concentration of employee owners, 84.51%[72] only have between 1% and 6% of shares held by employees. According to information provided by the latest PEPPER report[73], this situation can be explained by the fact that in many European countries financial participation emerged only in the mid-1980s and for new member states in Eastern Europe it is still a "recent phenomenon".

Nonetheless, in regard to the proposed hypotheses, the prevalence of low concentration levels among all investigated identities is conspicuous. If a higher concentration were optimal, a different distribution would be expected. This reasoning follows the argument by Demsetz and Villalonga[74], which states that if there was a systematic relation between ownership structure and firm performance, non-optimal dispersed structures should not survive.

3.4. Data Analysis

3.4.1. Multiple Linear Regression

The goal is to test the relation between ownership concentration of individual owner identities and firm performance. To test the proposed hypotheses, multiple linear regression is used. The model is constructed as follows:

$$\text{Performance}_i = \alpha + \beta_1 \text{ Concentration}_{is} + \sum \gamma' X_i + \sum \lambda_{1,t} \text{ Year}_t + \sum \lambda_{2,j} (\text{Country} \times \text{Industry})_j + u_i + \varepsilon_{it}, \quad (1)$$

where Performance represents (log of) Tobin's Q and ROA, $i = 1, \ldots, N_s$ is the firm, $s =$ Management, Family, Employee, Block is the owner identity, $\text{Concentration}_{is}$ represents the (log of) fraction of shares of an owner type[75] and X_i refers to a vector of firm-specific control variables which include (log of) the number of employees, (log of) leverage, and R&D intensity. Moreover, $t = 1, \ldots, 9$ is the year of observation and $j = 1, \ldots, 347$ is an interaction of the country and industry of observation. Besides the unobservable random error term ε_{it}, the clustered error term u_i is used to account for correlation within each firm i.

[69] See Richter and Schrader (2017), p. 404.

[70] See for example Anderson et al. (2003); Haldar and Rao (2011); Yixiang (2011).

[71] See van Pottelsberghe (2020), p. 6.

[72] $453/(453 + 71 + 8 + 4) = 0.8451$.

[73] See Lowitzsch et al. (2009), pp. 207-209 and pp. 81-198.

[74] See Demsetz and Villalonga (2001), pp. 209f.

[75] Note that for management and employee owner the aggregated shares are used to calculate concentration, while for family and block owners the number of shares of the largest owner is used.

Table 1: Descriptive Statistics for Selected Numeric Variables

Variable	N	Mean	25th pctl.	Median	75th pctl.	SD
Management Concentration	18,768	6.983	0.009	0.176	2.22	16.464
Family Concentration	18,898	5.269	0	0.02	4.07	12.04
Employee Concentration	18,767	1.014	0	0.213	1.003	2.263
Blockholder Concentration	18,898	35.066	13.16	29.5	51.48	25.177
Tobin's Q	17,611	1.65	1.024	1.296	1.813	1.123
ROA	18,250	0.072	0.008	0.043	0.095	0.409
Number of Employees	18,915	12.339	0.765	2.545	8.743	29.274
Market Capitalization	18,895	2,869.45	161.3	439.095	1,656.97	7,817.60
R&D Intensity	7,462	0.305	0.005	0.021	0.068	1.558
Leverage	18,248	0.234	0.08	0.217	0.346	0.182

Concentration Levels in Per Cent, Number of Employees in Thousands, Market Capitalization in Millions

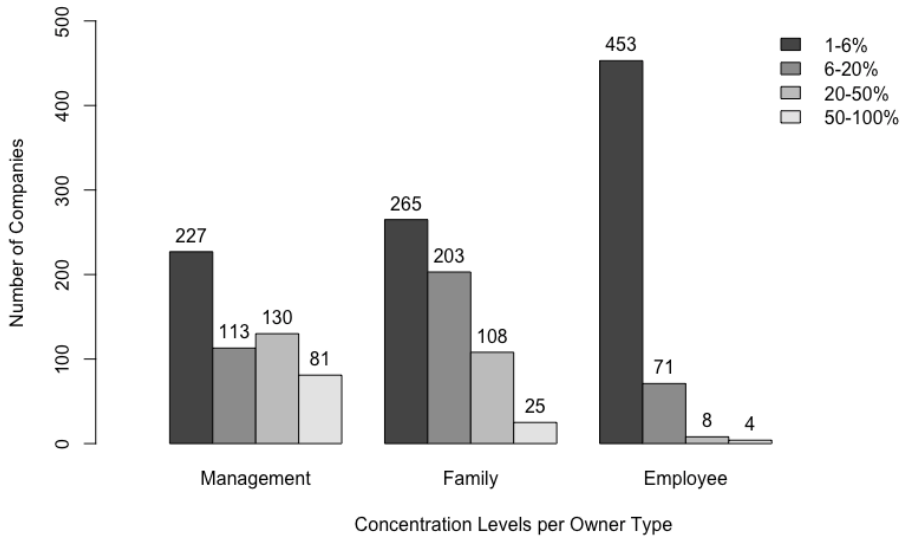


Figure 1: Number of Companies per Concentration Level per Owner in 2016

3.4.2. Multiple Linear Regression using Categorical Predictors

As referenced in section 3.2.2, there is reason to question the assumption of a linear relationship between the concentration of an owner type and firm performance. For this reason, a second model is used:

$$\text{Performance}_i = \alpha + \sum \beta' C_{isq} + \sum \gamma' X_i + \sum \lambda_{1,t} \text{Year}_t + \sum \lambda_{2,j} (\text{Country} \times \text{Industry})_j + u_i + \varepsilon_{it}, \tag{2}$$

where C_{isq} is a set of mutually exclusive dummy variables for each concentration level q = significant, strategic, determining, controlling of an owner s and a firm i. For instance, when the management's aggregated fraction of shares is 40%, $C_{(}$is,determining) is 1 and all other dummies are 0. It should be noted that C_{isq} has been constructed using $\text{Concentration}_{\text{is}}$. All other variables remain unchanged.

The goal of this model is to break the continuous concentration variable into four categorical predictors. Doing so allows to compare each individual concentration level to the baseline group in which an owner is not at all or not significantly owning company. If the investigated relationship is indeed non-linear, different coefficients and significance levels are expected for each of these dummy variables.

4. Results

4.1. Management Ownership and Firm Performance

Table 2 shows estimates for equation 1 in which (log of) Tobin's Q is regressed on the continuous log-transformed fraction of shares of management owners. Control variables are included stepwise in the order of their expected importance, which has been drawn from existing literature (see section 3.2.3). The adjusted R^2 increases for every model and all control variables are significant at the 1%-level, highlighting their relevance. Furthermore, the predicted coefficients

for the control variables are consistent with similar studies.[76] 'Number of Employees' and 'Leverage' have negative signs and do not show unexpected effect sizes. For instance, using 'Model 4' a 1% increase in the debt ratio decreases Tobin's Q by 0.37%.[77] Further contributing to the explanatory value of the models, the F-test is highly significant for every model.

The model meets the first three assumptions needed for consistent regression results using ordinary least squares, while the fourth untestable assumption is commented (see Appendix B for detailed explanation).

Management owner concentration has a positive and significant effect on Tobin's Q in 'Model 1' up to 'Model 3'. However, although this effect is rather small: Using 'Model 3', an increase from 20% to 30% in the relative number of shares held by managers improves a company's Tobin's Q by 1.11%.[78]

In 'Model 4', the effect is not significant. An explanation for this is the sharp decline in the number of observations (from 17,523 to 7,316) when the variable 'R&D Intensity' is introduced because it excludes a substantial number of firms that changed their ownership structure and hence, a lot of variance. Noting that the magnitude remains unchanged, it can be expected that a more exhaustive data set would lead to significant results. To sum up, Table 2 provides evidence in support with hypothesis 1.

Table 3 presents regression results when ROA is used as a dependent variable. It stands out that all estimates are highly significant. Although the results would support hypothesis 1, Table 3 raises doubts as to whether ROA is an applicable proxy for firm performance in this study. A potential explanation for the odd results is the data source. This is supported by looking at how ROA is expected to behave in relation to Tobin's Q. As Richter and Schrader[79] point out, ROA and Tobin's Q tend to be highly correlated. However, for the data sample used in this study, both variables turn out to be not correlated (Pearson correlation coefficient of 0.113). Regressing ROA also on the dummy variables again yields highly significant results. As, however, the sample size and the variance are small for high levels of concentration (see section 3.3) it is improbable to observe such a

significant effect. As a last remark, investigating the distribution of the residuals, a clear deviation from a normal distribution can be seen (see Appendix C). Applying the ROA models to other owner types, yield similar significant results. Although they are not conflicting in terms of direction when compared to Tobin's Q, they are questionable in terms of validity and are therefore not reported for the rest of this thesis.

Estimates for regressing the log of Tobin's Q on categorical predictors (equation 2) are presented in Table 4. It can be seen that results for ownership levels between 6 and 20 per cent are consistently significant across the first three models. However, their magnitude decreased distinctly. The effect can be interpreted as follows: Switching from no or non-significant levels of management ownership to a strategic level predicts a 6.61% [80] increase in the geometric mean of Tobin's Q, according to 'Model 11'.

However, this effect does not hold for all concentration levels, indicating that the relation between owner concentration and firm performance is indeed non-linear. Management ownership at a determining degree is significant at the 10% level when controlling for firm size and debt ratio ('Model 11'). However, higher owner concentrations are not significant and are only half as large in terms of the magnitude. When 'R&D Intensity' is introduced, the magnitude remains robust at all concentration levels, although not significant. Following the argument stated above, it is likely that the lack of significance is the result of the sharp decline in the number of observations.

Using the first three models, it can be concluded that the impact of management owners is largest when they are concentrated at a strategic level, and it becomes smaller when further increasing their concentration. Hence, according to all models, hypothesis 1 has to be rejected in the sense that the highest possible managing owner concentration would yield the highest possible performance. Nonetheless, the significance levels and the magnitudes indicate that the relation follows a non-linear curve that has a maximum somewhere between 6 and 50 per cent. It can be concluded that an increase in the relative number of management shares is favourable up to a certain degree.

Due to the fact that the focus of this study lies on the complete model including all control variables and that the change in results when fixed effects are introduced is similar across all owner identities, in the following, regression results are reported more concisely.

4.2. Family Ownership and Firm Performance

Table 5 presents estimates for equation 1 ('Model 13' and 'Model 14') and equation 2 ('Model 15' and 'Modell 16'). In all models, the log of Tobin's Q is used as a dependent variable.

Using the proportion of family owner shares as a continuous variable to measure concentration ('Model 13' and 'Model 14'), in both models, the effect on firm performance is not significant and very small: A 1% increase in 'Family Concentration' predicts an increase of 0.006%[81] in Tobin's Q.

Breaking down this variable into four dummies ('Model 15' and 'Model 16'), again reveals that there is a non-linear relationship. According to 'Model 15', compared to a company in which a family owns less than 1% of all outstanding shares, a company with significant family ownership is predicted to only have a 0.50%[82] higher Tobin's Q. This value

[76] See for example Aguilera and Crespi-Cladera (2016); Barontini and Caprio (2006); Demsetz and Villalonga (2001); Faleye et al. (2006).

[77] $1 - 1.01^{-0.374} = 0.0037$.

[78] $(1 + 0.3/0.2)^{0.012} = 1.0111$.

[79] See Richter and Schrader (2017), p. 403.

[80] $\exp(0.064) = 1.0661$.

[81] $1.01^{0.006} = 1,00006$.

[82] $\exp(0.005) = 1.0050$.

Table 2: Regression Results of log of Tobin's Q on Management Ownership Concentration

Variable	Model 1	Model 2	Model 3	Model 4
Management Concentration (log)	0.020***	0.021***	0.012**	0.012
	(0.005)	(0.005)	(0.005)	(0.008)
Number of Employees (log)			-0.020***	-0.022***
			(0.003)	(0.006)
Leverage (log)			-0.297***	-0.374***
			(0.044)	(0.076)
R&D Intensity				0.032***
				(0.008)
Constant	0.898***	0.956***	1.116***	1.012***
	(0.007)	(0.006)	(0.026)	(0.047)
Year FE	No	Yes	Yes	Yes
Country × Industry FE	No	Yes	Yes	Yes
N	17,595	17,534	17,523	7,316
F-statistic	122.322***	17.048***	18.871***	14.756***
Adjusted R^2	0.007	0.328	0.353	0.404

Dependent Variable: Log of Tobin's Q. Standard Errors are clustered on firm level.
***p <0.01, **p <0.05, *p <0.1

Table 3: Regression Results of Return on Assets on Management Ownership Concentration

Variable	Model 5	Model 6	Model 7	Model 8
Management Concentration (log)	0.004	0.016***	0.027***	0.025***
	(0.004)	(0.005)	(0.005)	(0.008)
Number of Employees (log)			0.040***	0.031***
			(0.005)	(0.006)
Leverage (log)			-0.407***	-0.566***
			(0.047)	(0.087)
R&D Intensity				-0.075***
				(0.01)
Constant	0.069***	0.159***	-0.123***	0.253***
	(0.009)	(0.007)	(0.034)	(0.043)
Year FE	No	Yes	Yes	Yes
Country × Industry FE	No	Yes	Yes	Yes
N	18,197	18,132	18,121	7,442
F-statistic	3.137*	14.090***	16.161***	17.100***
Adjusted R^2	0.0001	0.278	0.31	0.439

Dependent Variable: ROA. Standard Errors are clustered on firm level.
***p <0.01, **p <0.05, *p <0.1

increases to 3.05%[83] in case of strategic ownership. However, it decreases again when the family share proportion is controlling (- 0.50%) or determining (2.33%). Moreover, 'Strategic Ownership' is the only level showing a significant effect. These results change to only a small extent when also controlling for 'R&D Intensity'. The p-value for the dummy variables increase, but likely due to the decline in the number of observations.

[83] $\exp(0.030) = 1.0305$.

To conclude, 'Strategic Ownership' is the concentration related to the strongest firm performance. Hence, hypothesis 2 can only partly be accepted as higher family owner concentration is related to higher firm performance only up to a certain degree.

4.3. Employee Ownership and Firm Performance

Table 6 provides estimates for equation 1 ('Model 17' and 'Model 18') and equation 2 ('Model 19' and 'Model 20') when employees are the identity of observation. Again, all models

Table 4: Regression Results of Tobin's Q on Management Ownership Concentration (Categorical Predictors)

Variable	Model 9	Model 10	Model 11	Model 12
Significant Ownership	0.056***	0.032**	0.007	0.001
(≥ 1% & < 6%)	(0.016)	(0.014)	(0.013)	(0.023)
Strategic Ownership	0.121***	0.094***	0.064***	0.059
(≥ 6% & < 20%)	(0.027)	(0.024)	(0.022)	(0.039)
Determining Ownership	0.051**	0.070***	0.041*	0.04
(≥ 20% & < 50%)	(0.022)	(0.022)	(0.022)	(0.04)
Controlling Ownership	0.026	0.048*	0.021	0.03
(≥ 50%)	(0.027)	(0.026)	(0.026)	(0.053)
Number of Employees (log)			-0.020***	-0.022***
			(0.003)	(0.006)
Leverage (log)			-0.295***	-0.373***
			(-0.044)	(0.075)
R&D Intensity				0.032***
				(0.008)
Constant	0.896***	0.955***	1.116***	1.017***
	(0.007)	(0.006)	(0.026)	(0.047)
Year FE	No	Yes	Yes	Yes
Country × Industry FE	No	Yes	Yes	Yes
N	17,595	17,534	17,523	7,316
F-statistic	53.253***	17.050***	18.844***	14.666***
Adjusted R^2	0.012	0.33	0.354	0.404

Dependent Variable: Log of Tobin's Q. Standard Errors are clustered on firm level.
***p < 0.01, **p < 0.05, *p < 0.1

present a significant F-test and adjusted R^2 between 35% and 41% and thus acceptable explanatory power.

Using the fraction of shares that are held by employees as an independent variable (log-transformed), results are highly significant ('Model 17' and 'Model 18'). This also holds true when controlling for 'R&D Intensity', which has not been the case for other owner types. Applying 'Model 18', an increase from 0.5% to 1% in the employee share proportion would result in an expected increase of 6.93%[84] in Tobin's Q. Hence, 'Model 17' and 'Model 18' imply that hypothesis 3 can be accepted.

Using dummy variables to approximate the relationship's non-linearity, this impact remains constant for significant and strategic employee owners. For example, according to 'Model 20' switching from companies with non-significant to companies with significant employee owner concentration, a 6.82%[85] increase in the geometric mean of Tobin's Q is expected. When comparing non-significant with strategic ownership levels, this number is 11.07%[86].

Employee owners with determining or controlling concentration are not significantly related to firm performance. Apart from that, they show a negative effect, indicating that companies which are held to more than 20% by employees perform worse than companies that do not issue any or only a few shares to their employees.

To summarise, employee share ownership and firm performance are only positively related when the ownership concentration of employees does not exceed the strategic level. Thus, hypothesis 3 can only party be accepted.

4.4. Ownership Concentration and Firm Performance

To test the hypothesis that firms having a concentrated ownership structure perform better than comparable firms with a dispersed ownership structure, three further models are introduced. First, Tobin's Q (log-transformed) is regressed on the fraction of shares the largest shareholder is owning ('Model 21'). 'Model 22' uses a dummy variable to indicate whether a company is concentrated according to a 25% threshold. 'Model 23' repeats this step but with a 50% threshold. Estimates are presented in Table 7.

The adjusted R^2 and F-statistic, as well as all control variables, remain almost unchanged across all three models. The concentration variables differ only marginally from zero and are not significant. Although the coefficient for concentrated firms at a 50% threshold is smaller than the coefficient at the 25% threshold, no considerable differences can be observed.

To conclude, firms having a concentrated ownership structure are not related to superior firm performance. Therefore, hypothesis 4 is rejected.

[84] $(1 + 0.01/0.005)^{0.061} = 1.0693$.
[85] $\exp(0.066) = 1.0682$.
[86] $\exp(0.105) = 1.1107$.

Table 5: Regression Results of Tobin's Q on Family Ownership Concentration

Variable	Model 13	Model 14	Model 15	Model 16
Family Concentration (log)	0.006	0.006		
	(0.004)	(0.007)		
Significant Ownership			0.005	0.015
($\geq$ 1% & < 6%)			(0.012)	(0.02)
Strategic Ownership			0.030**	0.035
($\geq$ 6% & < 20%)			(0.015)	(0.024)
Determining Ownership			-0.005	-0.018
($\geq$ 20% & < 50%)			(0.019)	(0.03)
Controlling Ownership			0.023	0.022
($\geq$ 50%)			(0.034)	(0.068)
Number of Employees (log)	-0.021***	-0.023***	-0.021***	-0.022***
	(0.004)	(0.006)	(0.003)	(0.006)
Leverage (log)	-0.300***	-0.380***	-0.300***	-0.375***
	(0.043)	(0.076)	(0.043)	(0.077)
R&D Intensity		0.032***		0.031***
		(0.008)		(0.008)
Constant	1.122***	1.026***	1.123***	1.022***
	(0.027)	(0.047)	(0.028)	(0.047)
Year FE	Yes	Yes	Yes	Yes
Country × Industry FE	Yes	Yes	Yes	Yes
N	17,528	7,314	17,528	7,314
F-statistic	18.776***	14.704***	18.710***	14.630***
Adjusted R^2	0.352	0.403	0.353	0.404

Dependent Variable: Log of Tobin's Q. Standard Errors are clustered on firm level.
***p < 0.01, **p < 0.05, *p < 0.1

5. Discussion

5.1. Hypothesis 1

One finding of using equation 2 to analyse the relationship between management share concentration and firm performance is that it is not linear. For this reason, the positive and significant coefficients estimated in Table 2 (equation 1) cannot be used as a valid criterion to evaluate the first hypothesis. Observing that the dummy variables 'Strategic Ownership' and 'Determining Ownership' remain significant across 'Model 9' to 'Model 11' and that their magnitudes remain stable across all models, it has been concluded that ownership at such levels reliably shows a positive effect on firm performance. However, this conclusion may be premature without further investigation of potential weaknesses of the model.

One concern that should be addressed is the following: The majority of firms in the total sample do not have any or only have less than 1% of their total shares held by managers.[87] The performance of such firms, however, does not have to be curbed as other beneficial owners can be in place. For example, a company with a less than significant proportion of management shares can have a high relative firm performance that is caused by a large fraction of family owners. Therefore, comparing different management owner concentrations to a baseline in which owner-managers may not be represented at all, is detrimental to the model's significance. For this reason, 'Model 11' and 'Model 12' have been retested using a filtered data set that exclusively contains firms with an aggregated management share proportion bigger than zero (see Appendix A.1). The results show little deviation from the initial models. All dummy variables have positive coefficients in both models ('Model 11.b' and 'Model 12.b'), and significant, strategic and determining levels are significant in 'Model 11.b'. Hence increasing the fraction of shares held by managers from non-significant to strategic levels yields a 6.50%[88] increase in Tobin's Q and a 4.39%[89] increase when shifting to determining levels.

To summarise, the initial models used supported the hypothesis that a higher concentration of management owners leads to higher firm performance up to strategic and determining levels. This section further investigated one potential weakness of these models and proposed a modification to ac-

[87] 1 minus all firms that have a management owner concentration of more than 1% divided by all firms in the sample $= 1 - (227 + 113 + 130 + 81)/2000 = 0.7245 = 72.45\%$.

[88] $\exp(0.063) = 1.0650$.

[89] $\exp(0.043) = 1.0439$.

Table 6: Regression Results of Tobin's Q on Employee Ownership Concentration

Variable	Model 17	Model 18	Model 19	Model 20
Employee Concentration (log)	0.049***	0.061***		
	(0.011)	(0.016)		
Significant Ownership			0.065***	0.066***
(≥ 1% & < 6%)			(0.012)	(0.016)
Strategic Ownership			0.064*	0.105*
(≥ 6% & < 20%)			(0.034)	(0.059)
Determining Ownership			-0.009	-0.023
(≥ 20% & < 50%)			(0.037)	(0.045)
Controlling Ownership			-0.036	-0.038
(≥ 50%)			(0.063)	(0.061)
Number of Employees (log)	-0.026***	-0.029***	-0.026***	-0.028***
	(0.004)	(0.006)	(0.003)	(0.005)
Leverage (log)	-0.290***	-0.350***	-0.291***	-0.353***
	(0.043)	(0.076)	(0.044)	(0.076)
R&D Intensity		0.032***		0.032***
		(0.008)		(0.008)
Constant	1.146***	1.017***	1.158***	1.007***
	(0.026)	(0.046)	(0.026)	(0.046)
Year FE	Yes	Yes	Yes	Yes
Country × Industry FE	Yes	Yes	Yes	Yes
N	17,522	7,315	17,522	7,315
F-statistic	19.128***	15.054***	19.094***	14.943***
Adjusted R^2	0.357	0.409	0.358	0.409

Dependent Variable: Log of Tobin's Q. Standard Errors are clustered on firm level.
***p < 0.01, **p < 0.05, *p < 0.1

Table 7: Regression Results of Tobin's Q on Blockholder Concentration

Variable	Model 21	Model 22	Model 23
Blockholder	-0.0002		
Concentration	(0.0002)		
Blockholder (≥ 25%)		-0.004	
		(0.01)	
Blockholder (≥ 50%)			-0.00004
			(0.01)
Number of Employees (log)	-0.022***	-0.022***	-0.022***
	(0.003)	(0.003)	(0.006)
Leverage (log)	-0.302***	-0.301***	-0.301***
	(0.044)	(0.044)	(0.044)
Constant	1.137***	1.136***	1.133***
	(0.026)	(0.027)	(0.025)
Year FE	Yes	Yes	Yes
Country × Industry FE	Yes	Yes	Yes
N	17,528	17,528	17,528
F-statistic	18.755***	18.750***	18.748***
Adjusted R^2	0.352	0.352	0.352

Dependent Variable: Log of Tobin's Q. Standard Errors are clustered on firm level.
***p < 0.01, **p < 0.05, *p < 0.1

count for it. The results reinforce the conclusion that higher management concentrations are beneficial to a certain degree. A more in-depth examination to determine the position of this optimal level may be the topic of future research.

5.2. Hypothesis 2

The relation between family ownership and firm performance, again, turned out to be non-linear and emphasized the importance of equation 2 as a valid model. The results presented in section 4.2 support hypothesis 2 to the extent that firms having a strategic family ownership structure have a statistically significant higher Tobin's Q than firms with no or no significant family owners.

Nevertheless, this model suffers from the same interaction problem as the management model: The effect of low concentrations of family owners may be suppressed by the impact of larger, potentially well-performing owners. To address this concern, 'Model 15' and 'Model 16' have been retested using a filtered dataset that excludes firms without any family owners (see Appendix A.2). The results do not change, which highlights their robustness.

Comparing the magnitudes of the estimates for the concentration dummies with each other, an interesting pattern emerges. For instance, using 'Model 15', significant ownership does not lead to a noticeable effect in firm performance (0.005), strategic ownership leads to a significant positive effect (0.030), for determining ownership, the effect is not noticeable (-0.005), and for controlling ownership, it rises again (0.023). Thus, the relation indicates to have more than one optimal concentration level. This would mean that the simple bell-shaped curve, which has been proposed by Thomsen and Pedersen[90], is in fact more complex. Such a two-peak relation has been discovered by Morck, Schleifer and Vishny[91]. Although their finding was related to management owners, the argumentation can be applied to family owners as well. The initial increase in the effect on Tobin's Q can be explained by the increasing incentives of families to increase the value of their rising stakes. The special value of the families is demonstrated: synergies are used, information is shared, and resources are gathered.[92] After the first peak, however, entrenchment effects start to set in that exceed the benefits of the family owners. After a certain limit, however, there are no more limits for owners in a positive sense and a pure convergence-of-interest effect sets in, whereby the success of the company increases again.

It is questionable whether this effect actually arises from the interaction of entrenchment and convergence-of-interest effects and not from irregularities in the data set. The number of companies with a family concentration of more than 50% is low and last but not least the high p-values disprove a significant effect. Therefore, it is refrained from fully accepting the second hypothesis. The result remains that an increased concentration of family owners is only to some extent related to better performance. Hypothesis 2 is therefore only partially accepted.

5.3. Hypothesis 3

According to the results presented in 4.3, employee owner concentration and firm performance are positively related as long as the concentration levels do not exceed the strategic level. Hypothesis 3 is, therefore, partly accepted. This section will further explain why the results are significant for low concentration levels but not significant for higher concentration levels.

Using the decomposed dummy variables, it can be noted that employee ownership above a 20% threshold does not lead to better firm performance when compared to firms having no or less than significant levels of employee ownership. However, the sample size for firms representing such high concentration levels is far from sufficient to draw any meaningful conclusion. For example, for the year 2016, the sample contains only four companies, that are held to more than 50% by their employees. One possible explanation for this is that financial participation by employees, especially to a large extent, is not yet established (see section 3.3). The small number of examples that prove corporate success with controlling employee owners may be a self-fulfilling prophecy. The average capital held by employees has doubled in the last ten years, and there is currently no sign of this trend weakening.[93] Hence, it may take some time before an empirically valuable statement can be made about the impact of high employee concentrations on company performance. On the other hand, there are already some arguments against a good performance of very high concentrations. As already explained, the increasing interest of employees in the financial situation of the company can lead to a slowdown in decision-making processes and a decline in the productivity of the company.[94] An investigation of this hypothesis would be an interesting topic for future research, especially if the owner structure is measured by owner involvement and not by company shares.

If one looks at why employees are involved in companies in the first place (see section 2.2.3), it becomes clear that an important reason is to increase employee motivation, which is supposed to improve company performance. It is very likely, however, that even a low level of participation is sufficient to achieve this goal. The advantages that employees bring with them relate mainly to their knowledge around the product. This value cannot be compared with the value of access to capital and networks by, e.g. family owners. Moreover, some firms may introduce employee ownership only to materialize tax incentives, for which low concentrations are sufficient.

In summary, it can also be theoretically concluded that after a certain level, more employee ownership should not lead to further improvement in company performance.

[90] See Thomsen and Pedersen (2000).
[91] See Morck et al. (1988), pp. 301-302.
[92] See Silva and Majluf (2008), p. 613.
[93] Survey of Employee Ownership in European Countries 2017.
[94] See for example Park et al. (2004); Richter and Schrader (2017).

In order to test the robustness of the models used and to support the preliminary conclusion that employees have a positive influence on the company up to a strategic level, four further models were tested (see Appendix A.3). As with the other owners, the dummy models were tested with a filtered data set containing only records in which a company has at least a single employee owner ('Model 19.b' and 'Model 20.b'). The results show no noticeable difference, which confirms the validity of the models.

In order to examine hypothesis 3 further, the linear relationship was also tested at low concentration levels ('Model 17.b' and 'Model 18.b'). The goal was to refute the argument that the relation within low concentrations may not be positive. For this purpose, log of Tobin's Q was again regressed to the continuous variable 'Employee Concentration' (log-transformed). The difference is that the data set was previously filtered and contains only observations with an employee concentration between 0% and 6%. The result is highly significant at the 1% level and indicates that an increase in employee owner concentration is associated with better firm performance as long as no strategic level is exceeded.

In conclusion, the presumption is confirmed, and hypothesis 3 is accepted with one constraint.

5.4. Hypothesis 4

Hypothesis 4 looked at the relation between concentrated ownership structures in general and firm performance. Results for the continuous and both categorical models were insignificant, and hypothesis 4 was rejected.

This result is unexpected in that it contradicts numerous studies in the literature.[95] But it is also in line with the widely cited argument that underperforming ownership structures should not survive in the long run.[96]

It should be noted that the variable 'Blockholder Concentration' contains various owner identities besides managers, families and employees. The fact that evidence in favour of managers, families and employees has been found in this study highlights the value of these owners. They bring in special benefits that compensate their disadvantages up to a certain degree.

5.5. Limitations

5.5.1. Measuring Ownership

This study uses the fraction of shares of an owner (aggregated or by an individual) to measure the ownership structure of a company. The implicit assumption of this measure is that the proportion of shares directly translates to the possibility that an owner can realize his interests and improve firm performance. However, the number of shares is not necessarily equivalent to the power of an owner. Even owners with a high fraction of shares can be abused by owners with a small number of shares.[97] For instance, families may be able to exert a disproportionally strong influence even at low levels of concentration, while employees may face a lack of experience that would allow them to fully exploit their shares.

For this reason, the fraction of shares may not be an accurate proxy for ownership structure and measures that account for the involvement of an owner should be used. However, such data has not been available for this study and papers using different measures are not known.

5.5.2. Interaction Effects

It is a common approach in the literature to look at individual owner types in isolation, and most studies focus on a single owner identity. This, however, does not alter the fact that interaction effects can play an important role in explaining the impact of a single owner on firm performance. The existence of multiple different owner identities in a hybrid ownership structure gives rise to potential conflicts, in which case, the objectives of the dominant identity will prevail. For example, family owners may be sensitive to the existence of other owners.

As the initial models used in this study fail to address these effects, the models have been retested using a filtered dataset that requires the owner of observation to be present in all sample firms. But still, a labour stake of, e.g., 20% may not be of explanatory value when 50% of shares are held by managers. To account for this problem, a more rigorous approach has been chosen by Faleye, Mehrotra and Morck[98], who retested their results using a subset in which the owner of observation has to be the largest shareholder in all companies. This approach, however, is not applicable in this study, as the sample size of such a subset would be too small. A different solution would be the use of private firms that are held by a single owner type. Again, this is not applicable here due to data availability.

5.5.3. Endogeneity

The models used in this study fulfil the underlying assumptions that are needed for ordinary least squares regression to produce consistent estimates (See Appendix B for detailed description). The fourth assumption is empirically not testable, but the following can be noted.

Three potential sources have been identified, according to which this study would suffer from omitted variable bias. First, there exists the notion that ownership structure is the product of decisions made by current and future shareholders, e.g. the decision to offer new shares.[99] It includes the complexity of different interests that differ among owner identities. This concern is partly addressed in this study by modelling ownership structure as a set of different owner types to account for their difference in interests, analysed in separate regressions. The second source of endogeneity may be the implementation of the ownership structure. This issue is especially relevant for employee owners who often

[95] See for example Kapopoulos and Lazaretou (2007); Lins et al. (2013); Thomsen and Pedersen (2000); Yixiang (2011).

[96] See Demsetz (1983); Demsetz and Villalonga (2001).

[97] See Demsetz and Villalonga (2001), p. 215.

[98] See Faleye et al. (2006), p. 20.

[99] See Demsetz (1983), pp. 384-386.

become an owner in the course of ESOPs or other participation plans. The effectiveness of employee ownership depends on whether it is established in a collectivistic or in an individualistic society.[100] This issue is addressed by controlling for industry-country fixed effects.

Another omitted variable is the importance of experience and network effects that are gained when being a controlling owner for a certain time. According to Pedersen and Thomsen[101], as soon as one becomes owner, this owner will gain experience and build a network that will strengthen his position. For this reason, the ownership structure is the self-reinforcing result of different histories, cultures and trends in economic development. In Germany, for example, banks are often important blockholders of a company, as they used to play a significant supporting role during industrialisation.[102] To account for this issue, this study controls for industry-country fixed effects. However, following this notion, it can be argued that an optimal ownership structure has been reinforced and does not change much over time. Therefore, firms having a suboptimal ownership structure that changes over time affect this analysis.

In regard to simultaneity bias, it can be plausibly argued that either higher ownership concentration causes higher firm performance or that higher firm performance causes a higher ownership concentration. To further discuss the latter, it can be argued that ownership structure is the endogenous outcome of compensation contracting processes.[103] For example, if managers expect that the company in which they work will grow in value, their incentive to demand equity compensation is particularly high, and the ownership structure of the firm is adjusted.

As stated above the explanatory variables used in this study may be correlated with the unobservable, time-invariant variables of the error term, which is referred to as fixed effects.[104] To control for fixed effects, year and country-industry dummy variables have been introduced. Error-terms are clustered on firm-level to account for potential problems of autocorrelation and non-independence. To conclude, endogeneity concerns have been addressed using multiple approaches and are found to be minimised.

5.5.4. Data

The raw data used for this study show some weaknesses. Excluding owner concentrations of over 100% and companies without employees are just some of the filtering steps taken to correct these errors. A particularly important error, possibly due to reporting, is erratic changes in owner concentration. For example, in 52 family businesses, a family owner with more than 50% of all shares suddenly had no shares at all in the following year, but one or two years later almost the same number of shares as before. The same error happens for 120 firms with the dummy that indicates the 20% threshold. This is problematic in that it erroneously results in many major changes in the ownership structure, but hardly leads to any change in Tobin's Q. As a result, the impact of high concentration levels may be represented worse than it actually is. For this reason, the regressions were repeated for all models, with such jumps being corrected by the arithmetic mean of the adjacent values for the fraction of shares of an owner.

6. Conclusion

The relationship between ownership structure and firm performance has been the subject of a long-running discussion. However, whether this connection is positive, negative or even significant at all is still undecided in the literature. Researches have argued and found empirical evidence for every viewpoint. This quantitative study finds that (a) a higher concentration of management, family and employee owners is to some extent related to higher company performance, (b) this relationship follows a non-linear pattern, and (c) a concentrated ownership structure, ignoring the identity of the owner, does not influence firm performance.

The results show that when describing the ownership structure of a company, the identity of the owners must be taken into account. This is because different owner types have different interests and risk preferences. Ultimately, each owner brings certain benefits (e.g. financial and network access or information on product and process-specific optimization potential) but also disadvantages (e.g. entrenchment or inefficiency in the decision-making process). The optimal concentration of an owner is at the level where the advantages of an owner outweigh its disadvantages.

In this study, 2,120 companies from 31 European countries were examined using two different multiple linear regression models – one that uses a continuous variable to measure the fraction of shares held and one that divides this variable into five categorical concentration levels (non-significant, significant, strategic, determining and controlling). Using Tobin's Q as the primary performance measure, it is shown for management, family and employee owners, respectively, that companies that have such an owner at a strategic concentration level perform significantly better than comparable companies in which the owner is not or not significantly represented. Although significant, the effect is quite small, and practitioners must decide whether a change in the ownership structure yields a satisfactory high improvement in performance.

Future literature can make further interesting contributions to this. On the one hand, it is interesting to investigate more closely which factors are relevant for determining the optimal concentration levels. It can be assumed that every company must find an optimal ownership structure for itself. Given the possible measurement error, it is also interesting to conduct this study by including another measure for the ownership structure, which also takes involvement into account.

[100] See Caramelli and Briole (2007), pp. 297-300.
[101] See Pedersen and Thomsen (1997), pp. 761f.
[102] See Roe (1993), p. 1971, p. 1929.
[103] See Cho (1998), pp. 105-106.
[104] See Roberts and Whited (2012), p. 76.

References

Agnblad, J., Berglöf, E., Högfeldt, P., & Svancar, H. (2001). Ownership and control in sweden: Strong owners, weak minorities, and social control. *The control of corporate Europe*, 228-258.

Aguilera, R. V., & Crespi-Cladera, R. (2016). Global corporate governance: On the relevance of firms' ownership structure. *Journal of World Business*, *51*(1), 50-57.

Aguilera, R. V., & Jackson, G. (2003). The cross-national diversity of corporate governance: Dimensions and determinants. *Academy of management Review*, *28*(3), 447-465.

Anderson, R. C., Mansi, S. A., & Reeb, D. M. (2003). Founding family ownership and the agency cost of debt. *Journal of Financial economics*, *68*(2), 263-285.

Anderson, R. C., & Reeb, D. M. (2003). Founding-family ownership and firm performance: evidence from the s&p 500. *The journal of finance*, *58*(3), 1301-1328.

Ang, J. S., Cole, R. A., & Lin, J. W. (2000). Agency costs and ownership structure. *the Journal of Finance*, *55*(1), 81-106.

Barontini, R., & Caprio, L. (2006). The effect of family control on firm value and performance: Evidence from continental europe. *European Financial Management*, *12*(5), 689-723.

Berle, A., & Means, G. (1932). The modern corporation and private property (mac-millan, new york, ny).

Blasi, J., Freeman, R., & Kruse, D. (2016). Do broad-based employee ownership, profit sharing and stock options help the best firms do even better? *British Journal of Industrial Relations*, *54*(1), 55-82.

Caramelli, M., & Briole, A. (2007). Employee stock ownership and job attitudes: Does culture matter? *Human Resource Management Review*, *17*, 290-304. doi: 10.1016/j.hrmr.2007.07.002

Cho, M.-H. (1998). Ownership structure, investment, and the corporate value: an empirical analysis. *Journal of financial Economics*, *47*(1), 103-121.

Claessens, S., Djankov, S., & Lang, L. H. (2000). The separation of ownership and control in east asian corporations. *Journal of financial Economics*, *58*(1-2), 81-112.

Core, J. E., & Larcker, D. F. (2002). Performance consequences of mandatory increases in executive stock ownership. *Journal of Financial Economics*, *64*(3), 317-340.

David, P., O'Brien, J. P., Yoshikawa, T., & Delios, A. (2010). Do shareholders or stakeholders appropriate the rents from corporate diversification? the influence or ownership structure. *The Academy of Management Journal*, *53*(3), 636-654. Retrieved from www.jstor.org/stable/25684340

Demsetz, H. (1983). The structure of ownership and the theory of the firm. *The Journal of Law & Economics*, *26*(2), 375-390. Retrieved from www.jstor.org/stable/725108

Demsetz, H., & Villalonga, B. (2001). Ownership structure and corporate performance. *Journal of corporate finance*, *7*(3), 209-233.

Dow, G. K., & Putterman, L. (2000). Why capital suppliers (usually) hire workers: What we know and what we need to know. *Journal of Economic Behavior & Organization*, *43*(3), 319-336.

EFES. (2019). *Database of employee share ownership in european companies* (Vols. January 10th, 2020) [Web Page]. Retrieved from http://www.efesonline.org/Database%20of%20employee%20ownership/2019/Database%20of%20Employee%20Ownership%20in%20European%20Companies%20-%202019.pdf(January15,2020)

Faleye, O., Mehrotra, V., & Morck, R. (2006). When labor has a voice in corporate governance. *Journal of financial and quantitative analysis*, *41*(3), 489-510.

Fama, & French, K. R. (n.d.). *Detail for 48 industry portfolios.* Retrieved from https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/det_48_ind_port.html(January15,2020)

Freeman, R., Kruse, D., & Blasi, J. (2008). Worker responses to shirking under shared capitalism.

Haldar, A., & Rao, S. N. (2011). Empirical study on ownership structure and firm performance. *Indian Journal of Corporate Governance*, *4*(2), 27-34.

James, H. S. (1999). Owner as manager, extended horizons and the family firm. *International Journal of the Economics of Business*, *6*(1), 41-55. Retrieved from https://doi.org/10.1080/13571519984304 doi: 10.1080/13571519984304

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. Retrieved from http://www.sciencedirect.com/science/article/pii/0304405X7690026X doi: https://doi.org/10.1016/0304-405X(76)90026-X

Kaarsemaker, E., Pendleton, A., & Poutsma, E. (2009). Employee share ownership plans: A review.

Kapopoulos, P., & Lazaretou, S. (2007). Corporate ownership structure and firm performance: evidence from greek firms. *Corporate Governance: An International Review*, *15*(2), 144-158.

Kim, E. H., & Ouimet, P. (2014). Broad-based employee stock ownership: Motives and outcomes. *The Journal of Finance*, *69*(3), 1273-1319.

Kim, K. Y., & Patel, P. C. (2017). Employee ownership and firm performance: A variance decomposition analysis of european firms. *Journal of Business Research*, *70*, 248-254.

Kurtulus, F. A., & Kruse, D. L. (2017). *How did employee ownership firms weather the last two recessions?: Employee ownership, employment stability, and firm survival: 1999-2011*. WE Upjohn Institute.

Lins, K. V., Volpin, P., & Wagner, H. F. (2013). Does family control matter? international evidence from the 2008–2009 financial crisis. *The Review of Financial Studies*, *26*(10), 2583-2619.

Livingston, D. T., & Henry, J. B. (1980). The effect of employee stock ownership plans on corporate profits. *Journal of Risk and Insurance*, 491-505.

Long, R. J. (1978). The effects of employee ownership on organizational identification, employee job attitudes, and organizational performance: A tentative framework and empirical findings. *Human Relations*, *31*(1), 29-48.

Lowitzsch, J., Hashi, I., Woodward, R., & (Eds.). (2009). *The pepper iv report: Benchmarking of employee participation in profits and enterprise results in the member and candidate countries of the european union* (Thesis).

Martínez, J. I., Stöhr, B. S., & Quiroga, B. F. (2007). Family ownership and firm performance: Evidence from public companies in chile. *Family Business Review*, *20*(2), 83-94. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1741-6248.2007.00087.x doi: 10.1111/j.1741-6248.2007.00087.x

Mazibuko, N. E., & Boshoff, C. (2003). Employee perceptions of share ownership schemes: an empirical study. *South African Journal of Business Management*, *34*(2), 31-44.

Morck, R., Shleifer, A., & Vishny, R. W. (1988). Management ownership and market valuation: An empirical analysis. *Journal of financial economics*, *20*, 293-315.

Morck, R., Wolfenzon, D., & Yeung, B. (2005). Corporate governance, economic entrenchment, and growth. *Journal of economic literature*, *43*(3), 655-720.

Oswald, S. L., & Jahera Jr, J. S. (1991). The influence of ownership on performance: An empirical study. *Strategic Management Journal*, *12*(4), 321-326.

Park, R., Kruse, D., & Sesil, J. (2004). *Does employee ownership enhance firm survival?* [Book]. Emerald Group Publishing Limited.

Pedersen, T., & Thomsen, S. (1997). European patterns of corporate ownership: A twelve-country study. *Journal of International Business Studies*, *28*(4), 759-778.

Pendleton, A. (2002). *Employee ownership, participation and governance: a study of esops in the uk*. Routledge.

Richter, A., & Schrader, S. (2017). Levels of employee share ownership and the performance of listed companies in europe. *British Journal of Industrial Relations*, *55*(2), 396-420.

Roberts, M. R., & Whited, T. M. (2012). Endogeneity in empirical corporate finance. *Simon School Working Paper No. FR 11-29*. doi: http://dx.doi.org/10.2139/ssrn.1748604

Roe, M. J. (1993). Some differences in corporate structure in germany, japan, and the united states. *The Yale Law Journal*, *102*(8), 1927-2003.

Shleifer, A., & Vishny, R. W. (1986). Large shareholders and corporate control. *Journal of Political Economy*, *94*(3), 461-488.

Silva, F., & Majluf, N. (2008). Does family ownership shape performance outcomes? *Journal of Business Research*, *61*(6), 609-614.

Thomsen, S., & Pedersen, T. (2000). Ownership structure and economic performance in the largest european companies. *Strategic Management Journal*, *21*(6), 689-705.

Thomsen, S., & Rose, C. (2004). Foundation ownership and financial performance: Do companies need owners? *European Journal of Law and Economics*, *18*(3), 343-364.

van Pottelsberghe, B. (2020). Europe's r&d: Missing the wrong targets?

Villalonga, B., & Amit, R. (2006). How do family ownership, control and management affect firm value? *Journal of financial Economics*, *80*(2), 385-417.

Xia, F., & Walker, G. (2015). How much does owner type matter for firm performance? manufacturing firms in c hina 1998–2007. *Strategic Management Journal*, *36*(4), 576-585.

Yixiang, Z. (2011). Empirical research on ownership structure and corporate performance based on the perspective of dynamic endogeneity. *Energy Procedia*, *5*, 1878-1884.

Junior Management Science 6(1) (2021) 100-148

Junior Management Science

journal homepage: www.jums.academy

The Impact of Taxation on Managerial Incentives and the Use of Bonus Banks

Der Einfluss der Besteuerung auf Managementanreize und die Nutzung von Bonusbanken

Daniel Dyck

Universität Bielefeld

Abstract

This study examines how taxes and tax code features influence a firm owner's use of a managerial incentive scheme known as bonus bank. Bonus banks have a unique property: part of the manager's earned bonus in one period is withheld and subsequently paid out only in case future period's firm goals are met. While prior literature on bonus banks either neglected taxes or took managerial incentives as given, I incorporate both aspects into a multi-period agency model with a risk neutral, limited liability agent. The model especially captures corporate income taxes, loss-offset restrictions as well as a manager's flat, period-dependent income tax rate. The results are twofold: First, compared to fixed remuneration contracts, the use of bonus banks is adversely affected by all investigated taxes and tax code features, except for the loss-offset restrictions. Second, compared to periodical bonus remuneration, bonus banks possess an income smoothing characteristic, weakening the negative effect loss-offset restrictions and income tax progressivity have on a firm owner's future value. Overall, the results complement the literature on the decision relevance of taxes when designing compensation contracts.

Zusammenfassung

Diese Studie untersucht, wie Steuern und steuerlichen Regelungen die Nutzung eines Managementanreizsystems, bekannt als Bonusbank, aus Unternehmenseignersicht beeinflussen. Bonusbanken haben eine spezifische Eigenschaft: ein Teil des von einem Manager in einem Jahr verdienten Bonus wird einbehalten und nachträglich nur bei Erreichung zukünftiger Unternehmensziele ausgezahlt. Während in der bisherigen Literatur zu Bonusbanken Steuern vernachlässigt oder Manageranreize als gegeben angenommen wurden, integriere ich beide Aspekte in ein mehrperiodiges Agency-Modell mit risikoneutralem, haftungsbeschränktem Agenten. Das Modell erfasst insbesondere Ertragsteuern und Verlustverrechnungsbeschränkungen auf Unternehmensebene sowie eine proportionale, periodenabhängige Einkommensteuer beim Manager. Die Ergebnisse zeigen zweierlei: Erstens sinkt die Bonusbanknutzung gegenüber dem Fixlohnvertrag durch alle untersuchten Steuern und Regelungen mit Ausnahme der Verrechnungsbeschränkungen. Zweitens besitzt die Bonusbank einen glättenden Einkommenseffekt, der die Endvermögensreduktion aus den Verrechnungsbeschränkungen und der Einkommensteuerprogression abschwächt. Insgesamt ergänzen die Resultate das Schrifttum zur Entscheidungsrelevanz von Steuern bei der Gestaltung von Vergütungsverträgen.

Keywords: Prinzipal-Agenten Theorie, Steuereffekte, Bonusbank, mehrperiodige Vertragsstruktur.

1. Motivation

Die Trennung von Eigentum und Kontrolle ist ein wesentliches Merkmal von großen Kapitalgesellschaften der heutigen Zeit.[1] Die mit dieser Trennung einhergehende Auslagerung der Unternehmensführung an ein spezialisiertes Management ist aber aus Sicht der Eigner nicht nur mit Vorteilen verbunden. Das Management handelt in seinem eigenen Interesse, sodass sich die Eigner mit potentiell unerwünschten Managemententscheidungen konfrontiert sehen. Diese Interessendivergenz begründet den klassischen Anreizkonflikt zwischen Managern und Anteilseignern. Der Anreizkon-

[1] Vgl. erstmals Berle und Means (1932) in ihrem Aufsatz zur "separation of ownership from control".

DOI: https://doi.org/10.5282/jums/v6i1pp100-148

flikt ist in der betriebswirtschaftlichen Praxis allgegenwärtig und erlangt aufgrund der Bedeutung von Managemententscheidungen eine große Aufmerksamkeit. Prendergast formuliert treffend: "Incentives are the essence of economics."[2]

Durch geeignete Anreizschemata wird versucht, das Verhalten des Managements zu beeinflussen und eine Interessenkongruenz herzustellen. Typische Vergütungselemente für Manager sind neben kurzfristigen Entgelten wie einem fixen Grundgehalt und variablen Bestandteilen wie jährliche Bonuszahlungen auch andere langfristig orientierte, häufig aktienbasierte Vergütungsinstrumente. Diese klassischerweise genutzten Vergütungselemente sind insbesondere vor und in der Wirtschaftskrise um das Jahr 2007/08 in die Kritik geraten, da einige Unternehmen trotz der schlechten wirtschaftlichen Lage zum Teil hohe Boni ausgeschüttet haben.[3] Angesichts der Bonusausschüttung in Milliardenhöhe der Deutschen Bank für das Geschäftsjahr 2017 trotz eines Jahresfehlbetrags von 700 Millionen Euro scheint die Kritik nicht an Aktualität zu verlieren.[4] Gleiches gilt für die UBS, bei der trotz eines starken Aktienkursrückgangs im Geschäftsjahr 2018 keine Kürzung der Vorstandsvergütungen zu verzeichnen ist.[5]

Vermehrt kommt vom Gesetzgeber die Forderung auf, Manager nicht nur an Gewinnen, sondern auch an Verlusten des Unternehmens zu beteiligen.[6] Diese Möglichkeit bietet die sogenannte "Bonusbank", bei der ein Teil des in einem Jahr verdienten Bonus erst in Abhängigkeit von der zukünftigen Unternehmensentwicklung ausgezahlt wird bzw. bei schlechter Entwicklung verloren geht. Über Bonusbankvereinbarungen wird der Versuch unternommen, die Partizipation am Unternehmenserfolg komplett an die Unternehmensentwicklung zu koppeln. Hinsichtlich der Führungskräftevergütungen wird dies in der Literatur als Rückkehr zu partnerschaftlichen Konzepten interpretiert,[7] die ursprünglich durch die Trennung von Eigentum und Kontrolle umgangen werden sollten.

Die konkrete Ausgestaltung und Wirkung von Anreizschemata wird analytisch in Prinzipal-Agenten-Modellen untersucht.[8] Trotz der Vielzahl an Modellerweiterungen ist der Einfluss der Besteuerung auf die Wirksamkeit der Anreizschemata lange Zeit unberücksichtigt geblieben, obwohl steuerliche Verzerrungseffekte in vielen anderen Bereichen nachgewiesen wurden.[9] Bereits in einem sehr einfachen Prinzipal-Agenten-Modell zeigt sich, dass pareto-effiziente Lösungen unter Berücksichtigung von Steuern die Steuerlast nicht zwangsläufig minimieren.[10] Der Einfluss der Besteuerung ist damit keineswegs trivial und muss folglich im Zusammenspiel mit mehreren Faktoren, im Besonderen der Zielfunktionen, Risikopräferenzen und Entscheidungsvariablen der Akteure berücksichtigt werden.[11]

Mittlerweile wurden charakteristische Elemente von Steuersystemen sowie diskutierte, steuerpolitische Maßnahmen in einperiodige Prinzipal-Agenten-Modelle integriert und in Bezug auf eine Vielzahl von anreiztheoretischen Fragestellungen untersucht.[12] Die Analyse von mehrperiodigen Vertragsbeziehungen unter Berücksichtigung steuerlicher Anreizwirkungen steht allerdings noch aus.[13] Insgesamt erweist sich die "Vernachlässigung von Steuern bei der Vertragsgestaltung [als] rechtfertigungsbedürftig."[14] Die Bonusbank wird hingegen in der Literatur zwar als Anreizinstrument intensiv diskutiert, eine formal-analytische Untersuchung wurde aber bisher einzig in dem Beitrag von Janocha (2014) vorgenommen. In seiner Analyse wird die Existenz von Steuern allerdings komplett ausgeblendet, was zu Fehlentscheidungen führen kann.

Ziel der Arbeit ist es, herauszufinden, welchen Einfluss ertragsteuerliche Regelungen auf die Anreizwirkung und Vorteilhaftigkeit einer Bonusbankvereinbarung im Rahmen eines Prinzipal-Agenten-Modells haben. Die Vorteilhaftigkeitsprüfung dieses Anreizsystems erfolgt aus Anteilseignersicht, wobei den Eignern als alternative Vergütungsform ein fixes Grundgehalt oder ein fester, jährlich ausgezahlter Bonus zur Verfügung steht. Insbesondere wird in der Arbeit herausgestellt, welche steuerrechtlichen Regelungen die Bonusbank und welche die alternativen Vergütungsformen begünstigen.

Zunächst wird dazu in Kapitel 2 die Problemdarstellung vertieft. Neben der Einordnung von Managern und Anteilseignern in den Unternehmenskontext werden die in der Theorie und Praxis diskutierten Anreizsysteme für Manager sowie die steuerrechtlichen Grundlagen vorgestellt. Abgeschlossen wird das Kapitel mit der Darstellung der grundlegenden Modellstruktur, den Annahmen und Begrifflichkeiten in der Prinzipal-Agenten-Theorie.

Darauf aufbauend wird in Kapitel 3 ein einfaches Prinzipal-Agenten-Modell entwickelt, mit dem zunächst grundsätzliche Steuerwirkungen bei beobachtbarer und unbeobachtbarer Handlung des Agenten aufgezeigt werden. Hierzu werden steuerliche Aspekte aus der einschlägigen Literatur in einen Modellrahmen mit risikoneutralen Akteuren integriert. Insgesamt wird dadurch ein in dieser Form noch nicht analysierter Rahmen geschaffen, der bereits in der Literatur identifizierte Steuerwirkungen bestätigt.

Den Kern der vorliegenden Arbeit bildet das Kapitel 4. Basierend auf dem Modell des vorherigen Kapitels wird zur Analyse von Bonusbankvereinbarungen ein mehrperiodiges Modell formuliert. Im ersten Schritt wird die Bonusbank unter Ausschluss von Verlusten auf Unternehmensebene analysiert. Das mehrperiodige Grundmodell zeigt bereits zentrale

[2] Prendergast (1999), S. 7.
[3] Vgl. Becker und Ostrowski (2012), S. 527f.
[4] Vgl. Schreiber (2018), Abrufdatum: 31.10.2019.
[5] Vgl. Steck (2019), Abrufdatum: 31.10.2019.
[6] Vgl. z. B. BT-Drucksache 16/13433 (2009), S. 10.
[7] Vgl. O'Byrne (2014), S. 29; Stewart (1991), S. 234f.
[8] Vgl. Winter (2001), S. 527f.
[9] Vgl. Hundsdoerfer, Kiesewetter und Sureth (2008), S. 102.
[10] Vgl. Fellingham und Wolfson (1985), S. 12f.
[11] Vgl. Wagner und Dirrigl (1980), S. 8.
[12] Bauer, Kourouxous und Krenn (2018) verschaffen einen Überblick über die bereits untersuchten Fragestellungen zu steuerlichen Einflüssen im Manager-Anteilseigner-Konflikt.
[13] Vgl. Niemann (2011), S. 19; Bauer et al. (2018), S. 72.
[14] Niemann (2007), S. 2.

Steuerwirkungen auf, während die Erweiterung mit periodenabhängiger Einkommensteuer einen vertiefenden Blick auf den Einfluss der tariflichen Einkommensteuer des Managers in Bezug auf die optimale Vertragswahl liefert. In einem zweiten Schritt wird der Einfluss von Verlustentstehungsmöglichkeiten auf Unternehmensebene untersucht. Es werden relevante Konstellationen identifiziert und steuerinduzierte Effekte von Verlustverrechnungsbeschränkungen herausgestellt. Im Anschluss erfolgt die kritische Würdigung der Modellannahmen, die die Robustheit der hergeleiteten Ergebnisse prüft.

Die Arbeit schließt mit einer Zusammenfassung der Ergebnisse und einem Ausblick auf Ergebnisverwendungsmöglichkeiten.

2. Ökonomische Problemdarstellung

In diesem Kapitel werden Manager und Anteilseigner als Anspruchsgruppen in das allgemeine Konzept der Corporate Governance eingeordnet. Anschließend werden die in der Theorie und Praxis klassischerweise diskutierten Vergütungsformen von Managern vorgestellt und kritisch auf ihre Wirksamkeit geprüft. Auf den Kritikpunkten aufbauend wird eine detaillierte Funktionsbeschreibung der Bonusbank vorgenommen. Zudem werden die steuerrechtlichen Grundlagen herausgestellt, die auf den Ebenen der Akteure für die vorliegende Arbeit von Bedeutung sind. Einen zentralen Schwerpunkt bildet die einkommensteuerliche Qualifikation von Vergütungen aus Bonusbankvereinbarungen beim Manager und bei der betrachteten Kapitalgesellschaft. Zuletzt werden zentrale Termini aus der Prinzipal-Agenten-Literatur eingeführt, die für den Analyseteil notwendig sind. Gleichzeitig findet eine weitere Eingrenzung des Untersuchungsgegenstandes statt.

2.1. Corporate Governance und der Konflikt zwischen Managern und Anteilseignern

Unter dem Begriff Corporate Governance wird die "Organisation der Leitung und Kontrolle in einem Unternehmen mit dem Ziel des Interessenausgleichs zwischen den verschiedenen Anspruchsgruppen"[15] verstanden. Die Anspruchsgruppen werden als Stakeholder, die Unternehmenseigner als Shareholder bezeichnet. Nach Becker gehören zu den internen Stakeholdern neben den Shareholdern auch die Führungskräfte und weiteren Mitarbeiter im Unternehmen. Externe Stakeholder werden auf der einen Seite in primäre Stakeholder wie Kunden, Lieferanten und Fremdkapitalgeber und auf der anderen Seite in sekundäre Stakeholder wie den Staat und die Gesellschaft differenziert.[16]

Prinzipal-Agenten-Beziehungen liegen im einfachsten Fall vor, wenn ein Akteur (der Prinzipal) einen anderen Akteur (den Agenten) mit einer Aufgabe betraut, wobei der Agent den Weisungen des Prinzipals Folge zu leisten hat.[17] Charakteristisch für diese Beziehung ist das Prinzip der Arbeitsteilung und die hierarchische Struktur. Winter hebt hervor, dass die Arbeitsteilung ein wesentliches und notwendiges Prinzip in sämtlichen Unternehmen mit einer gewissen Mindestgröße sei. Die Trennung von Eigentum über das Unternehmen und Kontrolle der Geschäftstätigkeit stelle eine spezielle Form dieses grundlegenden Prinzips in entwickelten Volkswirtschaften dar.[18] Offensichtlich ist die Arbeitsteilung mit Vorteilen verbunden, da Unternehmenseigner aufgrund "zeitliche[r], physische[r] und kognitive[r] Restriktionen"[19] die Entscheidungen nicht selbst treffen können. Dies gelte besonders für die immer komplexer werdenden Produktions- und Distributionsprozesse in Unternehmen.[20]

Im Mittelpunkt der Arbeit steht der Interessenkonflikt zwischen Shareholdern und Managern. Der Fokus auf diesen Konflikt ist zum einen mit der langfristig erheblichen Wirkung der Entscheidungen durch Führungskräfte zu begründen. Aufgrund des grundsätzlichen Charakters können diese Entscheidungen häufig nur unter hohen Kosten rückgängig gemacht werden und sind daher aus Sicht der Eigner von hohem Interesse.[21] Zum anderen steht die Analyse von Bonusbanken als Entgeltsystem im Zentrum des Analyseteils. Bonusbanken sind jedoch mit Blick auf die geringe Nutzung von langfristigen Anreizsystemen in untergeordneten Hierarchieebenen selten vorzufinden.[22]

Der Interessenkonflikt resultiert aus dem Eigeninteresse der Manager. Während Unternehmenseigner eine langfristige Unternehmenswertsteigerung anstreben, ziehen Manager aufgrund ihrer häufig beschränkten Dauer im Unternehmen kurzfristig profitablere Projekte vor, um ihren persönlichen Vorteil durch Boni oder den Aufbau einer Reputation zu ziehen.[23] Zudem können Unternehmenseigner die Manager häufig nicht für Fehlentwicklungen verantwortlich machen, selbst wenn diese durch eine schlechte Arbeitsleistung der Manager hervorgerufen wird.[24] Die hieraus verringerten Unternehmensrückflüsse sind der in dieser Arbeit betrachtete Teil der Agenturkosten.[25]

Eine Trennung von Eigentum und Kontrolle in Reinform, durch die der Interessenkonflikt entsteht, liegt lediglich in Kapitalgesellschaften vor. Da die GmbH sowie die AG die häufigsten Kapitalgesellschaftstypen sind, werden zur Einordnung die Anspruchsgruppen beider Rechtsformtypen dargestellt. Die Corporate Governance ist rechtlich überwiegend

[15] Witt (2001), S. 85.
[16] Vgl. Becker (2015), S. 107f.
[17] Vgl. Jensen und Meckling (1976), S. 308.
[18] Vgl. Winter (2001), S. 491.
[19] Jost (2001b), S. 1.
[20] Vgl. Fama und Jensen (1983), S. 332.
[21] Dies entspricht den Erkenntnissen der Pfadtheorie. Vgl. Becker (2015), S. 60ff.
[22] Vgl. Frenz, Schulz und Petry (2011), S. 58f.
[23] Vgl. Jensen und Murphy (1990), S. 226; Shleifer und Vishny (1997), S. 742ff.
[24] Vgl. Niemann (2011), S. 1.
[25] Zu den Agenturkosten zählen außerdem die Kosten der Überwachung des Agenten und die für den Agenten entstehenden Kosten durch Zusicherung von eignerkonformem Verhalten, z. B. durch regelmäßige Berichterstattung. Vgl. Jensen und Meckling (1976), S. 308.

im AktG und im GmbHG festgelegt, aber unterliegt Gestaltungsmöglichkeiten.[26] In einer AG werden die operativen und strategischen Entscheidungen durch den Vorstand, in einer GmbH durch die Geschäftsleitung getroffen.[27] Sie sind die primäre Unternehmungsleitung der jeweiligen Gesellschaft.[28] Der Aufsichtsrat einer AG dient im Wesentlichen der Kontrolle des Vorstandes. Bei Überschreitung von gewissen Größen ist auch bei der GmbH ein Aufsichtsrat verpflichtend. Die Unternehmenseigner sind bei einer AG die Aktionäre, die die Hauptversammlung bilden. In einer GmbH sind die Anteilseigner die Gesellschafter bzw. die Gesellschafterversammlung.

In der Arbeit steht der Begriff Manager repräsentativ für die primären Unternehmungsleitungsorgane der Kapitalgesellschaften. Der Begriff Anteilseigner wird stellvertretend für die Eigenkapitalgeber einer GmbH bzw. einer AG verwendet. Die obige Beschreibung einer Prinzipal-Agenten-Beziehung beschränkt sich auf *einen* Prinzipal und *einen* Agenten. Damit wird implizit unterstellt, dass die Personen innerhalb der Eigenkapitalgeber und der primären Unternehmungsleitungsorgane Gruppen mit gleichgerichteten Interessen und identischen persönlichen Umständen darstellen. Die persönlichen Umstände umfassen im Wesentlichen die Konsumpräferenzen und steuerliche Gegebenheiten, insbesondere die zugrunde liegenden Steuersätze. Insofern wird nachfolgend stets von *einem* Manager bzw. Anteilseigner gesprochen.

Auf die explizite Berücksichtigung des Aufsichtsrates als Akteur wird nachfolgend aus Vereinfachungsgründen verzichtet, obwohl dieser mit Ausnahme von kleinen GmbHs zwingend ist. In diesem Zusammenhang ist herauszustellen, dass in börsennotierten Unternehmen der Aufsichtsrat die Vergütungen des Vorstandes festlegt.[29] Der Prinzipal vereinigt folglich die Aktionäre und den Aufsichtsrat, wobei die Aktionäre schlussendlich die zentrale Rolle einnehmen.[30] Eine mögliche Interpretation dessen ist, dass trotz Überwachung durch den Aufsichtsrat die Handlungen der Manager nicht komplett kontrollierbar sind. Eine perfekte Kontrollstruktur im Unternehmen liegt somit nicht vor. Deshalb muss der Prinzipal weitere extrinsische Anreize liefern, um den Manager zu zielkonformem Verhalten, also einer hohen Arbeitsanstrengung, zu bewegen.

Da eine Bewertung der Vorteilhaftigkeit der Anreizsysteme aus Sicht der Unternehmenseigner erfolgen wird, liegt eine Fokussierung auf den Shareholder-Value Ansatz vor.[31] Im Sinne des Shareholder-Value Ansatzes werden Unternehmenseigner Steuern in ihre Entscheidungskalküle einbetten, um nachsteuerlich vorteilhafte Anreizsysteme zu nutzen. In dem nächsten Abschnitt werden die in der Praxis etablierten Anreizsysteme zur Abmilderung des Interessenkonflikts dargestellt und kritisch beleuchtet. Hierbei werden empirische Erkenntnisse aus börsennotierten Unternehmen und gesetzgeberische Vorschläge zu Anreizsystemen ebenfalls integriert. Ein spezieller Fokus liegt auf dem (neueren) Konzept der Bonusbank.

2.2. Managementvergütungen in der Praxis

2.2.1. Klassische Vergütungssysteme

Ein betriebliches Anreizsystem ist nach Wild (1973) die "Summe aller bewußt [!] gestalteten Arbeitsbedingungen, die bestimmte Verhaltensweisen (durch positive Anreize, Belohnungen etc.) verstärken, die Wahrscheinlichkeit des Auftretens anderer dagegen mindern (durch negative Anreize, Sanktionen etc.)"[32]. Ein Teil des Anreizsystems ist stets das Entgelt- bzw. Vergütungssystem[33]. Die beiden wichtigsten Bestandteile sind hierbei die fixe Grundvergütung und die variable Komponente. Die variable Komponente ist in der Regel abhängig von dem Erfolg der Gesamtunternehmung und lässt sich in variable Barvergütungen und aktienkursabhängige Komponenten differenzieren.[34] In Theorie und Praxis existiert die unstrittige Erkenntnis, dass die variablen Vergütungselemente die Leistung eines Managers hinsichtlich des Unternehmenserfolges fördern können.[35] Im theoretischen Diskurs steht außerdem die relative Vergütungshöhe, d. h. die Relation von fixer zu variabler Vergütung, nicht aber die absolute Vergütungshöhe im Mittelpunkt.[36]

Abbildung 1 veranschaulicht die in dieser Arbeit näher beschriebenen Anreizsysteme.
Die kursive Hervorhebung der aktienbasierten Anreizsysteme soll verdeutlichen, dass diese zwar im Rahmen des Abschnittes 2.2 betrachtet, nicht jedoch im Analyseteil untersucht werden. Jedes Anreizsystem einer Führungskraft beinhaltet typischerweise ein vereinbartes fixes Grundgehalt und einen Bonus in Abhängigkeit von rechnungslegungsbasierten Bemessungsgrundlagen. Für den Bonus einer Periode werden regelmäßig Obergrenzen und Untergrenzen festgelegt.[37] Typische Bemessungsgrundlagen sind der Gewinn bzw. Earning before interests and taxes (EBIT), residualgewinnbasierte Gewinngrößen wie der Economic Value Added oder der Umsatz. Sie werden regelmäßig um außergewöhnliche

[26] Vgl. Becker (2015), S. 132.
[27] Siehe § 76 Abs. 1 AktG und § 35 Abs. 1 GmbHG.
[28] Vgl. Becker (2007), S. 29ff.
[29] Siehe § 87 Abs. 1 AktG und § 87a Abs. 1 AktG-E. AktG-E steht für Entwurf eines Gesetzes zur Umsetzung der zweiten Aktionärsrechterichtlinie (ARUG II) vom 20. März 2019 (Regierungsentwurf). Vgl. BT-Drucksache 19/9739, S. 11f.
[30] Zur Problematik von nicht eignerkonformem Verhalten durch den Aufsichtsrat vgl. Bebchuk und Fried (2003), S. 73f.
[31] Dieser Ansatz geht auf Rappaport (1986) zurück.

[32] Wild (1973), S. 47.
[33] Vgl. Becker (1990), S. 12.
[34] Vgl. Döscher (2014), S. 15. In der Literatur hat sich der Begriff Bonus für die variable Barvergütung etabliert. In der Arbeit ist damit streng genommen eine Tantieme gemeint, da diese per Definition abhängig vom Erfolg der Gesamtunternehmung ist, während dies bei einem Bonus nicht der Fall sein muss. Der Einfachheit halber werden diese Begriffe als Synonyme verwendet.
[35] Vgl. bspw. Jensen und Murphy (1990), S. 226.
[36] Vgl. Winter (2001), S. 507. Gleichwohl stehen die absoluten Vergütungshöhen in der gesellschaftlichen Kritik. Sie werden aufgrund des Shareholder-Value-Ansatzes vernachlässigt.
[37] Vgl. Witzemann und Currle (2004), S. 632.

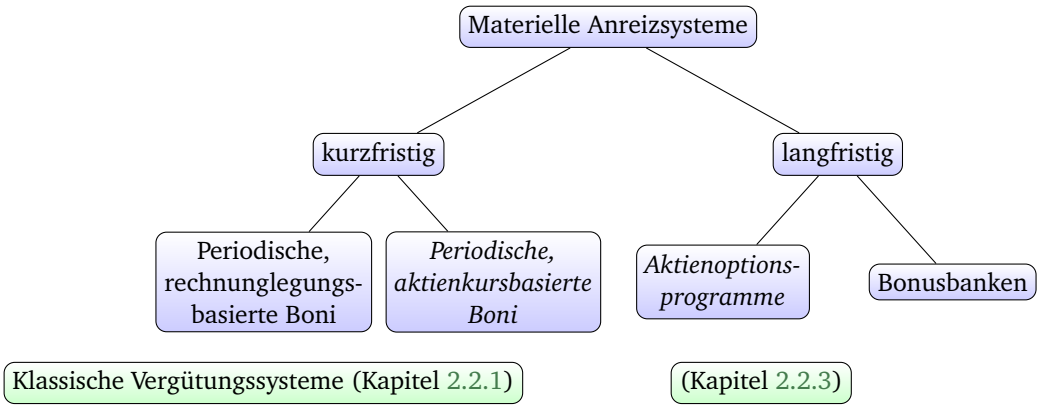


Abbildung 1: Systematisierung der in Kapitel 2.2 behandelten Anreizsysteme (Eigene Darstellung)

Geschäftsvorfälle bereinigt.[38] Zunächst empfahl der Gesetzgeber als Bemessungsgrundlage für Boni gewinnorientierte Größen in dem Bewusstsein, dass Umsatzgrößen leichter durch den Manager zu manipulieren seien.[39]

Das Grundgehalt stellt eine Mindestentlohnung und daher einen Mindestnutzen sicher, entfaltet aber keinen Leistungsanreiz. Damit gehört es nicht zu den Anreizsystemen. Der Bonus fördert hingegen die Anstrengung des Managers, sofern die Arbeitsanstrengung des Managers, die Höhe der Bemessungsgrundlage und der Erfolg der Gesamtunternehmung positiv korreliert sind.[40] Der Gesetzgeber fordert, dass alle variablen Bestandteile "in einem angemessenen Verhältnis zu den [...] Leistungen des Vorstandsmitgliedes"[41] stehen. Für den Unternehmenseigner ist dies derjenige Leistungseinsatz des Managers, bei dem unter Berücksichtigung von Anstrengungskosten dieser Leistung der Grenzeffekt der Anstrengung auf den (nachsteuerlichen) Unternehmenserfolg gerade mit der (vorsteuerlichen) Grenzentlohnung, die zu entrichten ist, ausgeglichen wird. Insgesamt gehört die variable Barvergütung zu den kurzfristig orientierten Anreizsystemen, wenn sie für ein Jahr vereinbart wird.[42]

Aktienbasierte Vergütungen werden in dieser Arbeit zu den klassischen Vergütungssystemen gezählt. Aufgrund der Vielfalt der Gestaltungsmöglichkeiten werden mit dem Bonus auf Basis des Börsen- bzw. Aktienkurses als Spezialfall der variablen Barvergütung und dem Aktienoptionsprogramm nur zwei zentrale Möglichkeiten dargestellt. Die grundsätzliche Idee hinter einer Vergütung mit dem Aktienkurs als Bemessungsgrundlage liegt aus Sicht des Gesetzgebers darin, dass in einem perfekten Kapitalmarkt der Aktienkurs den Wert des Unternehmens aus Sicht der Eigner widerspiegele. Wenn der Manager an der Aktienkursentwicklung beteiligt werde, so würde er letztlich die Ziele der Eigner übernehmen.[43] Hierbei wird die Maximierung des Börsenwertes häufig mit der Maximierung des Shareholder-Value gleichgesetzt.[44]

Aktienoptionsprogramme sind die am weitesten verbreitete, aktienbasierte Vergütungsform.[45] Die Aktienoption wird typischerweise als Call-Option festgelegt und beinhaltet das Recht des Managers, eine bestimmte Anzahl von Aktien zu einem festgelegten Ausübungspreis zu kaufen. Der Wert der variablen Entlohnung des Managers ist direkt an den Börsenkurs gekoppelt, da der Manager bei gestiegenem Börsenkurs in Höhe der Differenz zum Ausübungspreis einen Veräußerungsgewinn realisieren kann.[46] Ein Vorteil aus Unternehmenssicht ist hierbei, dass Aktienoptionen zu keinem Liquiditätsabfluss im Vergleich zur variablen Barvergütung auf Basis des Aktienkurses führen.[47] Um der Aktienoption eine langfristige Orientierung zu verleihen, schreibt der Gesetzgeber darüber hinaus Sperrfristen vor, innerhalb derer die Option nicht ausgeübt werden darf.[48] Im Gegensatz zu der variablen Barvergütung auf Basis des Aktienkurses werden Aktienoptionsprogramme zu den langfristigen Anreizsystemen gezählt.[49]

Abbildung 2 zeigt die Entwicklung der Vergütungsstruktur von CEOs der 50 größten Unternehmen in den USA von 1936 bis 2005.
Neben der steigenden absoluten Vergütungshöhe ist festzustellen, dass das Fixum und der Bonus auf Basis bestimmter operativer Zielgrößen bereits lange zur Anreizsetzung genutzt wird. Die absolute Höhe hat sich für diese Komponente im Zeitablauf kaum geändert, während die relative Höhe im Zeitablauf abnimmt. Dies liegt an der zunehmenden Wichtigkeit aktienbasierter Vergütungen und weiterer langfristiger Anreizsysteme seit 1950.[50] Zwischen 2000 und 2005 lag der Wert der Optionen in amerikanischen Unternehmen bei

[38] Vgl. Döscher (2014), S. 17.
[39] Vgl. BT-Drucksache 13/9712 (1998), S. 23.
[40] Außerdem muss für einen Leistungsanreiz erfüllt sein, dass der marginale Nutzen aus dem zusätzlichen Bonus die marginalen Anstrengungskosten des Managers deckt. Siehe Kapitel 2.4.
[41] § 87 Abs. 1 Satz 1 AktG.
[42] Vgl. Becker und Ostrowski (2012), S. 527f.
[43] Vgl. BT-Drucksache 13/9712 (1998), S. 23.
[44] Vgl. Witt (2001), S. 86. Dies ist aber kritisch zu sehen, siehe Kapitel 2.2.2.
[45] Vgl. Niemann und Simons (2003), S. 1.
[46] Vgl. Winter (2001), S. 522.
[47] Vgl. Arnold und Gillenkirch (2007), S. 77.
[48] Siehe § 193 Abs. 2 Nr. 4 AktG.
[49] Vgl. Becker und Ostrowski (2012), S. 528.
[50] LTIP steht für Long-Term Incentive Plan. Ein spezifischer LTIP, die Bonusbank, wird in Kapitel 2.2.3 beschrieben.

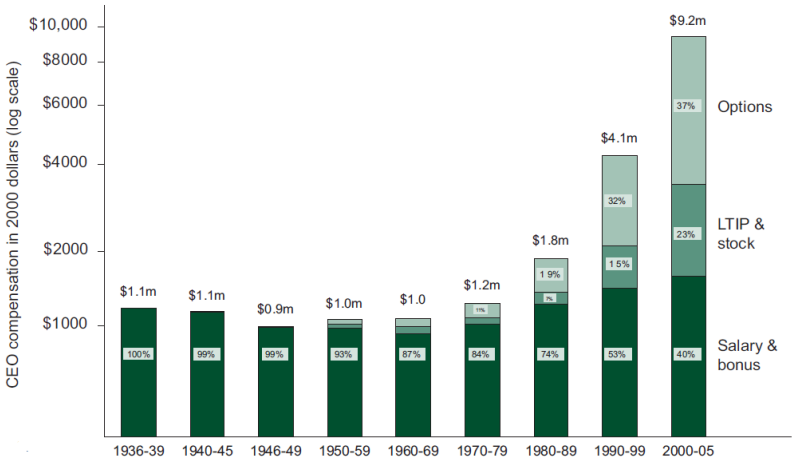


Abbildung 2: Struktur der CEO-Vergütung von 1936 bis 2005 in den USA (Frydman & Jenter, 2010, S. 80)

37% der gesamten variablen Vergütung. Werden die langfristigen Anreizsysteme und Belegschaftsaktien hinzu addiert, so übersteigen die langfristig orientierten die kurzfristig orientierten Vergütungsbestandteile wertmäßig.

Da es in Deutschland keine vergleichbare, langfristige Untersuchung gibt, wird diese amerikanische Studie zur Veranschaulichung genutzt. Es ist zu erwarten, dass es in deutschen Unternehmen einen ähnlichen Strukturverlauf gibt. Ein Unterschied ist jedoch, dass in Deutschland die aktienbasierte Vergütung erst seit 1990 und somit mit zeitlicher Verzögerung im Vergleich zu amerikanischen Unternehmen an Bedeutung gewonnen hat. Insbesondere mit dem Gesetz zur Kontrolle und Transparenz im Unternehmensbereich (KonTraG) ist jedoch ein signifikanter Anstieg zu verzeichnen gewesen.[51] Im Jahre 2017 nutzten 27 von 30 DAX30-Unternehmen aktienkursbasierte Vergütungen für ihre Vorstandsvorsitzenden, wobei der wertmäßige Anteil an der variablen Vergütung im Durchschnitt bei etwa 30% liegt. Der Anteil an kurz- und langfristig variablen Barvergütungen liegt bei 43%. Das Gehalt macht somit etwa 27% aus.[52] Die zeitliche Verzögerung der Entwicklung in Deutschland wird besonders im Vergleich zum Wert der aktienbasierten Vergütungen der 30 amerikanischen Unternehmen im Dow Jones Index deutlich. Hier beträgt der wertmäßige Anteil an den variablen Vergütungen im Durchschnitt knapp zwei Drittel, was ein Indiz für eine weitere Zunahme der Nutzung seit 2005 ist. Das Grundgehalt macht hier im Durchschnitt nur noch etwa 11% aus.[53]

[51]Vgl. Sanders und Tuschke (2007), S. 40f.

[52]Vgl. Friedl, Hölz, Beck, Heilmann und Pflugmann (2018), S. 5.

[53]Vgl. Friedl et al. (2018), S. 10. Die Studie von Frydman und Jenter (2010) kann aufgrund der Anzahl der Unternehmen und der unterschiedlichen Kategorisierung nur bedingt mit den Erkenntnissen von Friedl et al. (2018) verglichen werden. Würde man die kleineren Unternehmen in Frydman und Jenters Studie herausnehmen und um die unterschiedliche Kategorisierung der Vergütungsinstrumente bereinigen, so ergäbe sich vermutlich nur eine leicht geringere prozentuale Abweichung bei den aktienbasierten Vergütungsinstrumenten.

2.2.2. Kritik an klassischen Vergütungssystemen

Insgesamt ist innerhalb der klassischen Vergütungssysteme spätestens seit 1990 die Tendenz erkennbar, periodische, rechnungslegungsbasierte Barvergütungen in geringerem Maße zu nutzen. Diesbezüglich hat sich die Meinung des Gesetzgebers geändert. Er sieht nunmehr eine Leistungsmotivation ausschließlich durch diese Boni unabhängig von der Bemessungsgrundlage (d. h. auch auf Basis des Aktienkurses) kritisch, wenn die Bewertung stichtagsbezogen erfolgt.[54] Sie führt zu “fehlerhafte[n] Verhaltensanreize[n] [und] ist eine der Lehren der Finanzmarktkrise.”[55]

Zum einen liegt dies daran, dass die jeweiligen Bemessungsgrundlagen das Eignerziel der Unternehmenswertmaximierung nur verzerrt abbilden können.[56] Dieses Problem bleibt auch dann bestehen, wenn mehrere Kennzahlen zur Leistungsbeurteilung herangezogen werden.[57] Zum anderen bestehen Möglichkeiten der Bilanzpolitik, durch die der Manager die Höhe der Bonusbemessungsgrundlagen verändert, welche aber gleichzeitig Wertminderungen hervorrufen.[58] Dieses dysfunktionale Verhalten des Managers entsteht unter anderem durch die Bonusobergrenzen bzw. -untergrenzen (Caps bzw. Floors). Ist bspw. ein mit dem Prinzipal vereinbartes Gewinnziel für den Manager in einer Periode unerreichbar, so wird er Verluste in dieser Periode akkumulieren oder in dieser Periode ein geringes Anstrengungsniveau in dem Wissen wählen, dass er durch den Floor einen Mindestbonus erhält. Gleiches gilt für den umgekehrten Fall, in dem der Agent das Ziel in jedem Fall erreicht und er keinen Anreiz hat, die Zielgröße aufgrund des Caps weiter in die Höhe zu treiben.[59]

Neben den Tantiemen auf Basis jährlicher Kennzahlen

[54]Vgl. BT-Drucksache 16/13433 (2009), S. 1.

[55]BT-Drucksache 16/13433 (2009), S. 1.

[56]Vgl. Döscher (2014), S. 47.

[57]Vgl. Hostettler und Stern (2004), S. 170f.

[58]Vgl. Janocha (2014), S. 45ff.

[59]Vgl. Stewart (1991), S. 234.

ist der hohe Anteil der Aktienorientierung in der Kritik. Er wird als Mitverursacher der Finanzkrise 2007/08 gesehen, obwohl die empirischen Erkenntnisse hierzu nicht eindeutig sind.[60] Der zentrale Kritikpunkt besteht darin, dass die Aktienorientierung mit einer hohen Aufnahme von Risiken durch einige, insbesondere in der Finanzbranche tätige Manager einherging, die zu Lasten der Unternehmenseigner war.[61] Diese grundsätzliche Kritik an aktienbasierten Vergütungen schließt Industrieunternehmen indes nicht aus, da diese dort ebenfalls einen wesentlichen Bestandteil der Anreizstruktur ausmachen. Darüber hinaus wird bemängelt, dass die Kausalitätsrichtung zwischen der Gewährung von Aktienoptionen und dem Unternehmenserfolg nicht festzustellen ist. Es ist unklar, ob der Unternehmenserfolg durch die Nutzung von Aktienoptionen steigt oder ob die Nutzung von Aktienoptionen wegen einer erhöhten Erwartung des Unternehmenserfolges steigt.[62] Allerdings ist zu erwarten, dass die Gültigkeit der zweiten Kausalitätsrichtung umso wahrscheinlicher ist, je höher die Verhandlungsmacht des Managers ist.[63]

Um dieser fehlerhaften Anreizsetzung entgegenzuwirken, beschloss der Gesetzgeber das Gesetz zur Angemessenheit der Vorstandsvergütung (VorstAG). In § 87 Abs. 1 Satz 2 AktG wird gefordert, dass die "Vergütungsstruktur [..] bei börsennotierten Gesellschaften auf eine nachhaltige Unternehmensentwicklung auszurichten" ist. Des Weiteren sollen "[v]ariable Vergütungsbestandteile [...] eine mehrjährige Bemessungsgrundlage haben."[64] Ähnliche Empfehlungen, besonders im Hinblick auf die langfristige Orientierung, sind überdies im Deutschen Corporate Governance Kodex und in § 87a Abs. 1 AktG-E manifestiert.[65] Obwohl der Fokus auf börsennotierten Unternehmen liegt, so sollen diese Empfehlungen und Vorschriften analog in anderen Gesellschaftsformen wie der GmbH, sofern möglich, umgesetzt werden.[66]

Bei der konkreten Ausgestaltung der Anreizsysteme lässt der Gesetzgeber Gestaltungsspielräume. Allerdings hebt er den Nutzen von Bonus-Malus-Systemen hervor, die den Manager an negativen Entwicklungen des Unternehmens partizipieren lassen und eine Leistungsbetrachtung über mehrere Perioden vornehmen.[67] Die Idee, den Manager an negativen Wertentwicklungen teilhaben zu lassen, ist nicht neu, sondern spiegelt sich überdies in der zu Beginn des Kapitels 2.2.1 aufgezeigten Definition eines Anreizsystems von Wild (1973) wider. Die "Bonusbank" ist hierbei ein auszahlungsbasiertes Anreizsystem, welches viele der genannten Kritikpunkte berücksichtigt und zu einer langfristigen Motivation des Managers beitragen kann. Zudem ist die Nutzung im Gegensatz zu aktienbasierten Vergütungen ebenfalls in nicht börsennotierten Gesellschaften möglich.[68] Die Eckpunkte einer Bonusbank werden im kommenden Abschnitt vorgestellt.

2.2.3. Bonusbank als Entgeltsystem

Das Konzept der Bonusbank geht auf Stewart (1991) zurück. Historisch betrachtet entstand es als Kritik an klassischen Bonussystemen auf Basis rechnungslegungsbasierter Kennzahlen mit Caps und Floors und der Tatsache, dass Manager in der Praxis an positiven, nicht aber an negativen Unternehmensentwicklungen beteiligt wurden. Bereits im Jahr 2003 nutzen 29% der DAX100-Unternehmen eine Bonusbank.[69] Sie macht damit einen erheblichen Teil der Führungskräftevergütungen aus und ist von steigender Relevanz. Abbildung 3 verdeutlicht die zentrale Funktionsweise der Bonusbank im Vergleich zu klassischen Bonussystemen. Dabei wird davon ausgegangen, dass die Unternehmenseigner dem Manager eine Zielperformance vorgeben, welche im Regelfall eine (residual-)gewinnbasierte Größe zur Bemessungsgrundlage hat und dem Manager eine bestimmte Zielbonushöhe bringt.[70]

Das dysfunktionale Verhalten durch die Bonusbegrenzungen wird mit einer Bonusbank verhindert, da die Bonushöhe nach oben und unten offen ist.[71] Dadurch wird dem Manager stets ein Anreiz dazu gegeben, sich im Sinne des Unternehmens anzustrengen. Ist das Sollziel bspw. nicht zu erreichen, so wird er dazu bestrebt sein, zumindest einen (hohen) Malus zu verhindern. Im Idealfall sieht die Bonusbank bei einem negativen Zielerreichungsgrad eine Rückzahlungsverpflichtung von früher ausgeschütteten Boni oder Gehältern vor, um den Manager in eine eigentümerähnliche Lage zu bringen. Die direkte Beteiligung an Verlusten des Unternehmens durch Rückzahlungsverpflichtungen ist in der Praxis allerdings arbeitsrechtlich kaum durchzusetzen.[72] Eine Studie von Aders und Hebertinger bestätigt dies. Sie fanden heraus, dass nur 5% der DAX100- Unternehmen im Jahr 2003 Rückzahlungsverpflichtungen vereinbarten.[73] Deshalb wird der in einer Periode entstehende Malus in künftige Perioden vorgetragen und mit einem positiven Bonusbankguthaben aus künftigen Perioden verrechnet.[74] Dadurch wird eine partielle Verlustbeteiligung bewirkt und eine langfristige Sichtweise im Sinne des Unternehmens und des Gesetzgebers gefördert.[75]

Zudem wird aus Abbildung 3 ersichtlich, dass der bei positivem Zielerreichungsgrad erwirtschaftete Bonus, der in die Bonusbank eingestellt wird, nicht komplett in dieser Periode ausgezahlt wird. Lediglich ein konstanter Anteil zwischen 20% und 50% wird ausgeschüttet. Dieser Auszahlungsmodus

[60] Vgl. Benmelech, Kandel und Veronesi (2010), S. 1769f; Döscher (2014), S. 70f.
[61] Vgl. Bhagat und Bolton (2014), S. 314.
[62] Vgl. Core, Guay und Larcker (2003), S. 34f.
[63] Hierbei wird vom Managerial-Power-Ansatz gesprochen. Vgl. Bebchuk und Fried (2003), S. 72f.
[64] § 87 Abs. 1 Satz 3 AktG.
[65] Vgl. DCGK (2019), S. 30; BT-Drucksache 19/9739 (2019), S. 11.
[66] Vgl. BT-Drucksache 16/13433 (2009), S. 10.
[67] Vgl. BT-Drucksache 16/13433 (2009), S. 10.
[68] Vgl. Witzemann und Currle (2004), S. 631.
[69] Vgl. Aders und Hebertinger (2003), S. 38.
[70] Vgl. Hostettler und Stern (2004), S. 183; Witzemann und Currle (2004), S. 631f.
[71] Vgl. Stewart (1991), S. 233.
[72] Vgl. Witzemann und Currle (2004), S. 636.
[73] Vgl. Aders und Hebertinger (2003), S. 38.
[74] Vgl. Witzemann und Currle (2004), S. 636.
[75] Vgl. Witzemann und Currle (2004), S. 632.

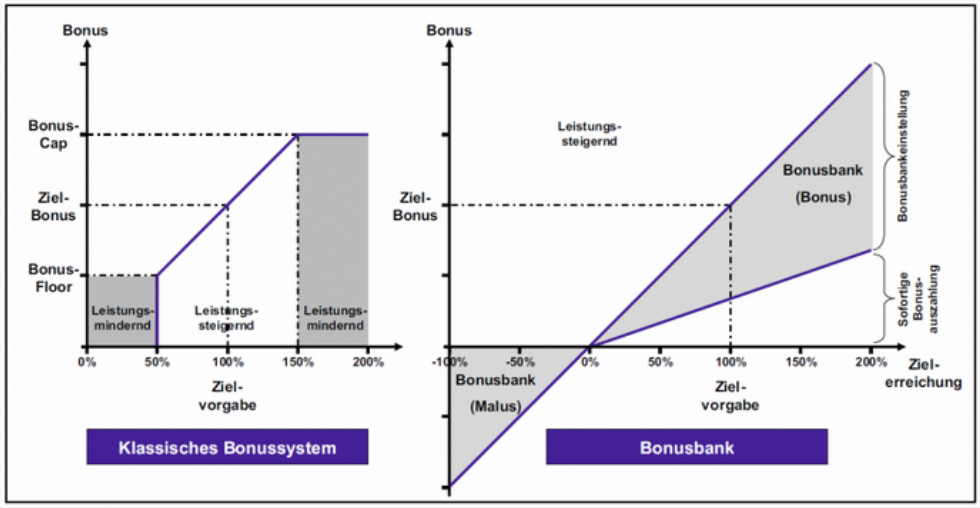


Abbildung 3: Funktionsvergleich von klassischen Bonussystemen gegenüber Bonusbanken (Witzemann & Currle, 2004, S. 633)

ohne "Schwelle" wird im Analyseteil in den Blick genommen. Im Gegensatz dazu wird bei einem Auszahlungsmodus mit "Schwelle" bis zu einer Zielerreichung in Höhe der Zielvorgabe der korrespondierende Bonus komplett ausbezahlt und darüber hinaus nur ein konstanter Anteil, der bei etwa 30% liegt.[76] Am Ende des Vereinbarungszeitraums, der in der Regel bei drei bis fünf Jahren liegt, wird der gesamte positive Kontostand an den Manager ausgeschüttet. Ist dieser Betrag negativ, so besteht die Möglichkeit, diesen in eine Folgevereinbarung zu übernehmen. Andernfalls entfällt der negative Saldo. In jedem Fall entfällt der negative Saldo, sofern der Manager das Unternehmen verlässt (bzw. verlassen muss).[77] Beendet der Manager das Arbeitsverhältnis vor dem Endzeitpunkt der Vereinbarung, so geht ein Teil oder typischerweise das komplette Guthaben verloren. Da ein Manager sein positives Guthaben nicht verlieren möchte, führt dies zu einer Bindung von leistungsstarken Managern.[78] Mit Blick auf den intensiven Wettbewerb um qualifizierte Führungskräfte ist dies ein weiterer Vorteil der Bonusbank.

Neben diesen Grundprinzipien kann zu Beginn der Bonusbankeinführung ein Startguthaben festgelegt werden. Das Startguthaben gibt dem Manager die Möglichkeit, Projekte durchzuführen, die kurzfristig zu Verlusten führen, aber langfristig profitabel sind. Der Einfluss eines Startguthabens wird gleichwohl vordergründig in Zielkongruenzmodellen wirksam, sodass es in dieser Arbeit nicht von Bedeutung ist.[79] Aufgrund des Kaufkraftverlustes durch die verzögerte Bonusauszahlung kann das nicht ausgeschüttete Guthaben verzinst werden. Häufig wird der Gesamtkapitalkostensatz des Unternehmens als Zinssatz zugrunde gelegt.[80] Als Alternative bietet sich ebenfalls der risikofreie Zins an. Die Nutzung des risikofreien Zinses kann damit gerechtfertigt werden, dass ex ante der nicht ausgeschüttete Anteil sicher feststeht und in den Folgeperioden ausschließlich eine mögliche Verlustbeteiligung unsicher ist.[81]

Zur Illustration der grundlegenden Funktionsweise ist in Tabelle 1 ein Zahlenbeispiel gegeben. In dem Beispiel wird von einer Ausschüttungsrate von 40% und einer Nominalverzinsung i von positivem (und negativem) Guthaben nach Auszahlung in Höhe von 10% ausgegangen. Der Vereinbarungszeitraum beträgt drei Jahre und der Manager verlässt das Unternehmen innerhalb dieses Zeitraums nicht. Zum Endzeitpunkt wird das gesamte Guthaben an den Manager ausgeschüttet. Die erste Zeile gibt die bereits bestimmten Bonushöhen abhängig von einem determinierten Unternehmensergebnis der Perioden $t = 1, 2$ und 3 an.
In der ersten Periode erwirtschaftet der Manager einen Bonus in Höhe von 100 GE, von dem lediglich 40 GE ausgeschüttet werden. In der zweiten Periode müsste der Manager bei direkter Ergebnisbeteiligung einen Malus von 50 GE hinnehmen. Dieser wird jedoch nicht fällig, da ein hoher Guthabensübertrag inklusive Verzinsung aus der Vorperiode vorliegt. In der Periode $t = 2$ wird deutlich, dass trotz des schlechten Periodenerfolges ein Bonus ausgezahlt wird. Die Bonusbank glättet folglich die Bonuszahlungen über den Vereinbarungszeitraum.[82] In der dritten Periode ist der Manager mit 70 GE am Unternehmenserfolg beteiligt. Zusammen mit dem Übertrag aus der Vorperiode wird das gesamte Bonusbankguthaben von 80,56 GE ausgezahlt.

[76] Vgl. Hostettler und Stern (2004), S. 179.
[77] Vgl. Crasselt und Fründ (2014), S. 166.
[78] Vgl. Stewart (1991), S. 241.
[79] Die Bonusbank entwickelte sich ursprünglich als Konzept zur Angleichung der Zielsetzungen von Manager und Anteilseigner ähnlich wie dies Residualgewinnkonzepte intendieren. Dabei geht es darum, dass der Manager die für das Unternehmen kapitalwertmaximierenden Projekte wählt. Vgl. Janocha (2014), S. 76, Fußnote 5.
[80] Vgl. Witzemann und Currle (2004), S. 635. Dieser wird zur Übereinstimmung mit dem Kapitalwertkriterium gewählt und ist eher für oben genannte Zielkongruenzmodelle entscheidend.
[81] Vgl. Crasselt und Fründ (2014), S. 166.
[82] Vgl. Stewart (1991), S. 240.

Periode	t=1	t=2	t=3
Anteilsmäßiger Bonus auf Basis des jährlichen Unternehmensgewinns	100	-50	70
Übertrag inklusive Verzinsung (i =10%)		66	10,56
Kontostand vor Auszahlung	100	16	80,56
Kontostand nach Auszahlung	60	9,6	0
Auszahlungsbetrag	40	6,4	80,56

Tabelle 1: Schematisches Beispiel für die Funktionsweise einer Bonusbank (In Anlehnung an Crasselt & Fründ, 2014, S. 167)

Die Kombination eines einzelnen Jahresziels mit einer mehrperiodigen Bonusbank sowie die partielle Verlustbeteiligung führen insgesamt zu einer kurz- und langfristigen Verhaltenswirkung, während traditionelle Bonussysteme lediglich kurzfristig ausgerichtet sind. Dabei ist die Bonusbank nicht auf eine bestimmte Bemessungsgrundlage beschränkt. Sie kann unabhängig von dieser zur langfristigen Leistungsmotivation des Managers genutzt werden.[83]

Steuerliche Vorschriften beeinflussen hierbei die Wirksamkeit aller Anreizsysteme. Im Mittelpunkt der Arbeit stehen steuerliche Wirkungen auf liquiditätswirksame Entgeltsysteme, d. h. insbesondere das Grundgehalt, der Bonus auf Basis rechnungslegungsbasierter Kennzahlen und die Bonusbank. Im nächsten Abschnitt werden die zentralen Vorschriften aus dem deutschen Steuerrecht dargestellt, die die Wirksamkeit dieser Vergütungssysteme auf den Ebenen der Akteure im Manager-Anteilseigner-Konflikt beeinflussen. Hierzu erfolgt darüber hinaus eine erste Prämissensetzung im Hinblick auf den Analyseteil, um den Fokus auf die relevanten Vorschriften zu lenken.

2.3. Steuerrechtliche Grundlagen

2.3.1. Auf Managerebene

Der konfessionslose Manager unterliegt annahmegemäß im Inland einer unbeschränkten Steuerpflicht gemäß § 1 Abs. 1 EStG. Ferner ist er kein Gesellschafter- Geschäftsführer, d. h. er ist nicht zugleich Anteilseigner der betrachteten Kapitalgesellschaft, in der er angestellt ist. Der Manager erhält für seine Tätigkeit eine Entlohnung. Unabhängig davon, ob diese Entlohnung als fixes Gehalt oder variable Entlohnung (jährliche Tantieme bzw. Bonusbank) festgelegt wird, stellen die Vergütungen des Agenten Einnahmen aus nichtselbstständiger Arbeit dar.[84] Der Zeitpunkt der Besteuerung der Einnahmen liegt in dem Kalenderjahr, "in dem sie dem Steuerpflichtigen zugeflossen sind."[85] Dies gilt ebenfalls für Bezüge wie Tantiemen, die "nicht als laufender Arbeitslohn gezahlt"[86] werden.

Wird der Führungskraft das jährliche Grundgehalt oder die Tantieme abhängig von dem Unternehmensergebnis der jeweiligen Periode komplett in der Verdienstperiode ausbezahlt, so ist der Zuflusszeitpunkt eindeutig. Entscheidend für den Zufluss ist nach ständiger Rechtsprechung der Übergang der "wirtschaftlichen Verfügungsmacht"[87] an den Arbeitnehmer. Auch bei einem verzögerten Auszahlungsplan wie der Bonusbankvereinbarung kann es abhängig von der Gestaltung dazu kommen, dass ein sofortiger Zufluss des einbehaltenen Anteils anzunehmen ist. Nach dem Urteil des pBFH ist dieser gegeben, falls der Manager eine Wahlmöglichkeit zur sofortigen Ausschüttung des einbehaltenen Anteils habe oder die Einbehaltung dem überwiegenden Interesse des Managers entspreche. Das Kriterium der wirtschaftlichen Verfügung beim Manager ist dann erfüllt. Eine Guthabensverzinsung des Bonusbankguthabens ist ebenfalls für die Festlegung des Zuflusszeitpunktes zu beachten. Dies gilt besonders, wenn der Arbeitnehmer als Fremdkapitalgeber identifiziert werde und Zinsen daher die Verfügungsmacht des Managers abbilden.[88] Werden allerdings Zinsen in angemessener Höhe gezahlt, sodass diese als "Entgelt für eine hinausgeschobene Fälligkeit"[89] qualifiziert werden, entspricht der Auszahlungszeitpunkt dem Zuflusszeitpunkt. In diesem Fall sind die anfallenden Zinsen als Einnahmen aus nichtselbständiger Arbeit zu versteuern.[90]

Bei der Bonusbank wird nicht von einem sofortigen Zufluss des einbehaltenen Anteils ausgegangen. Dies ist zum einen nicht erfüllt, da der einbehaltene Bonus leistungsabhängig ausgezahlt wird. Der Manager hat folglich kein Wahlrecht auf eine sofortige Ausschüttung. Zum anderen wird im Analyseteil unterstellt, dass der Manager einen sofortigen Bonus gegenüber einem späteren Bonus in gleicher Höhe vorzieht. Daher liegt die verzögerte Auszahlung ebenfalls nicht im Interesse des Managers. Eine Verzinsung wird im Analyseteil nicht explizit berücksichtigt. Sie spiegelt sich jedoch implizit in den Entlohnungszahlungen wider, die dem Manager aus einer Bonusbank zufließen. Insoweit erfüllt die Verzinsung das Kriterium der Angemessenheit.

Nach dem objektiven Nettoprinzip sind lediglich Nettoeinkünfte beim Manager steuerbar. Er kann grundsätzlich die mit den Einnahmen in Zusammenhang stehenden Werbungskosten bei der Einkünfteermittlung abziehen. Während die

[83] Vgl. Crasselt und Fründ (2014), S. 168; Janocha (2014), S. 30.
[84] Siehe § 19 Abs. 1 Satz 1 Nr. 1 EStG.
[85] § 11 Abs. 1 Satz 1 EStG.
[86] § 38a Abs. 1 Satz 3 EStG.
[87] BFH v. 14.05.1982, VI R 124/77, BStBl. II 1982, S. 472.
[88] Vgl. BFH v. 14.05.1982, VI R 124/77, BStBl. II 1982, S. 472.
[89] BFH v. 14.05.1982, VI R 124/77, BStBl. II 1982, S. 472.
[90] Vgl. Jasper (1995), S. 120.

tatsächlich angefallenen Werbungskosten durch die in der Analyse unterstellte Grenzbetrachtung außerhalb des Modells liegen, wird der Arbeitnehmerpauschbetrag vernachlässigt.[91] Ferner darf die am Ende des Kalenderjahres gezahlte Einkommensteuer nicht von den Einnahmen aus nichtselbständiger Arbeit als Werbungskosten abgezogen werden.[92] Damit sind die Einnahmen aus nichtselbständiger Arbeit zugleich die relevanten Einkünfte. Da Sonderausgaben und persönliche Umstände außen vor bleiben, entsprechen diese Einkünfte aus nichtselbständiger Arbeit dem zu versteuernden Einkommen (zvE) gemäß § 2 Abs. 5 EStG. Das zvE ist die Bemessungsgrundlage für die tarifliche Einkommensteuer des § 32a EStG, der die Entlohnungen unterworfen werden. Zusätzlich zur Einkommensteuer wird der Solidaritätszuschlag als Ergänzungsabgabe erhoben. Er beträgt 5,5% der festzusetzenden Einkommensteuer.[93]

Der Einkommensteuertarif ist in fünf Tarifzonen eingeteilt. Bis zu einem zvE in Höhe des Grundfreibetrags von € 9.168 fällt keine Einkommensteuer an. Übersteigt das zvE den Grundfreibetrag, wird ein Eingangsgrenzsteuersatz von 14% erhoben. Dieser steigt in der unteren Progressionszone (zvE bis € 14.254) und in der oberen Progressionszone (zvE bis € 55.960) bis zu einem Grenzsteuersatz von 42% an. Bei einem höheren zvE wird der Steuerpflichtige mit einem konstanten Spitzengrenzsteuersatz in Höhe von 42% belastet (Proportionalzone I). Ab einem zvE von € 265.327 greift schlussendlich die konstante Reichensteuerbelastung von 45% (Proportionalzone II).[94] Im Rahmen einer Grenzbetrachtung sind für den Manager stets die Grenzsteuersätze von Bedeutung. Er berücksichtigt bei seiner Entlohnung, inwieweit der zusätzlich verdiente (marginale) Bruttobetrag durch die marginale Steuerbelastung gemindert wird. Bei dem hier betrachteten Manager ist zu erwarten, dass der letzte verdiente Euro mit dem Spitzen- oder Reichensteuersatz belastet wird.

Die Anwendung der "normalen", tariflichen Einkommensteuer ist bei Zahlungen aus der Bonusbank allerdings nicht so offensichtlich wie bei einem Jahresgehalt oder der periodisch ausgeschütteten Tantieme. Insbesondere wird auf außerordentliche Einkünfte bei Vergütungen aus mehrjähriger Tätigkeit mit der Fünftelregelung ein begünstigter Steuersatz angewandt.[95] Die Voraussetzungen der Mehrjährigkeit sind bei einer Bonusbankvereinbarung erfüllt, sofern der Manager im Vereinbarungszeitraum nicht entlassen wird. Dann erstreckt sich die Tätigkeit über die geforderten zwei Veranlagungszeiträume und einen Zeitraum von mehr als einem Jahr.

Die Anwendung der begünstigten Steuerermittlung setzt grundsätzlich die Tatbestandsmerkmale der Außerordentlichkeit, der Zusammenballung von Einkünften und der Abgrenzbarkeit von laufenden Einkünften voraus.[96] Die Tatbestandsmerkmale sind bei Einkünften des § 19 EStG regelmäßig erfüllt, falls die Zusammenballung der Einkünfte auf "wirtschaftlich vernünftigen Gründen"[97] basiert. Das Merkmal der Außerordentlichkeit und der Abgrenzung zu laufenden Einkünften wird bei einer abhängigen Beschäftigung nicht zwangsläufig benötigt.[98] Sofern der Manager die Zielsetzungen der jeweiligen Perioden bei einer Bonusbankvereinbarung immer erreicht, kann es am Ende des Vereinbarungszeitraums zu einer Ausschüttung in signifikanter Höhe kommen, da sich Überträge aus den Vorperioden anhäufen. Insofern kann die Zusammenballung und ein daraus entstehender Progressionsnachteil gegeben sein. Dabei ist es nicht von Bedeutung, ob tatsächlich ein derartiger Nachteil beim Manager entsteht.[99]

Obwohl die Zusammenballung bei einer Bonusbank durch die langfristige Ausrichtung der Vergütung auf wirtschaftlich vernünftigen Gründen beruht,[100] sind nachgelagert ausgeschüttete, akkumulierte Guthabensbeträge aus Bonusbanken nicht ausnahmslos als außerordentliche Einkünfte zu qualifizieren. Eine begünstigte Besteuerung liegt vor, wenn bei einer Bonusbank der gesamte verdiente Bonus einbehalten und am Ende der Vereinbarung ausgezahlt wird.[101] Dies ist bei einer Auszahlung am Ende des Zeitraums eindeutig, da eine mehrjährige Zweckbestimmung der Entlohnung gegeben ist.[102]

Bei einer anteiligen Ausschüttung (siehe Kapitel 2.2.3) ist fraglich, ob eine derartige Trennung der Zahlungen in laufende Vergütungen und mehrjährige Tantieme möglich ist. Von Gerichten muss im Einzelfall entschieden werden, ob eine abgeltende Wirkung für vergangene Leistungen durch die verzögert ausgezahlten Boni beabsichtigt ist oder lediglich die Chance auf zukünftige Sondervergütungen eingeräumt wird.[103] Im ersten Fall werden die Zahlungen dem § 34 EStG, im zweiten Fall dem Regelsteuersatz des § 32a EStG unterworfen. Die Zweckbestimmung ergibt sich aus dem "Anlass der Zuwendung oder aus den übrigen Umständen"[104]. Wenn eine zentrale Funktion der Vereinbarung die Bindung von Mitarbeitern ist, so entspricht dies eher einer Zukunftsorientierung als einer Vergangenheitsorientierung.[105] Im Zweifel ist die Berechnungsart der nachträglich ausgezahlten Vergü-

[91] Siehe § 9 und § 9a Satz 1 Nr. 1 Buchstabe a EStG. Bei einer Führungskraft spielt der Pauschbetrag in Anbetracht der Vergütungshöhe eine untergeordnete Rolle.

[92] Siehe § 12 Nr. 3 EStG.

[93] Siehe § 1 Abs. 1 SolZG i. V. m. § 4 Satz 1 SolZG. Von der Darstellung etwaiger Befreiungen und Sonderregelungen beim Solidaritätszuschlag wird aufgrund der nur marginalen Auswirkungen abgesehen.

[94] Siehe § 32a Abs. 1 Satz 2 Nrn. 1-5 EStG.

[95] Siehe § 34 Abs. 2 Nr. 4 EStG i. V. m. § 34 Abs. 1 EStG.

[96] Vgl. Mellinghoff (2018), S. 1865ff, Rn. 6-13.

[97] R 34.4 Abs. 1 Satz 3 EStR.

[98] Siehe R 34.4 Abs. 1 Satz 3 EStR i. V. m. R 34.4 Abs. 2 EStR; dazu Mellinghoff (2018), S. 1876, Rn. 32.

[99] Vgl. BFH v. 15.10.2003 - XI R 17/02, BStBl. II 2004, S. 267; dazu Mellinghoff (2018), S. 1865f, Rn. 8.

[100] Wirtschaftlich vernünftige Gründe liegen nach ständiger Rechtsprechung lediglich bei willkürlichen, missbräuchlichen Konstruktionen nicht vor. Vgl. BFH-Urteil v. 07.05.2015, VI R 44/13, BStBl. II 2015, S. 892, Rn. 17.

[101] Vgl. FG Nürnberg, Urteil v. 10.05.2017, 3 K 1935/15, Rn. 4.

[102] Vgl. Mellinghoff (2018), S. 1876, Rn. 30.

[103] Vgl. FG Nürnberg, Urteil v. 10.05.2017, 3 K 1935/15, Rn. 34.

[104] FG Nürnberg, Urteil v. 10.05.2017, 3 K 1935/15, Rn. 32.

[105] Vgl. BFH-Urteil v. 18.03.1965, IV 116/64 U, BStBl. III, S. 290, Rn. 10.

tung maßgeblich.[106]

Insgesamt lässt sich feststellen, dass Bonusbankgestaltungen möglich sind, die sowohl zur Regelbesteuerung als auch zur Tarifbegünstigung führen. Dies wird in dem Urteil des Finanzgerichts Nürnberg besonders deutlich, weil dort eine steuerlich differenzierte Behandlung von Komponenten des zugrunde liegenden Bonusprogramms bestätigt wurde. Bei klassischen Bonusbankvereinbarungen ist jedoch von einer Regelbesteuerung des § 32a EStG auszugehen.

2.3.2. Auf Ebene der Kapitalgesellschaft

Die Kapitalgesellschaft, in der der Manager angestellt ist, hat seinen Sitz im Inland und ist daher unbeschränkt körperschaftsteuer- und gewerbesteuerpflichtig.[107] Zusätzlich zur Körperschaftsteuer wird der Solidaritätszuschlag erhoben.[108] Da die Kapitalgesellschaft ein Formkaufmann ist, ergibt sich hieraus die handelsrechtliche Buchführungspflicht.[109] Diese schlägt ebenfalls auf das Steuerrecht durch.[110] Die steuerliche Gewinnermittlung erfolgt durch einen Betriebsvermögensvergleich, wobei nach dem Maßgeblichkeitsprinzip Wertansätze dem Grunde und der Höhe nach aus der Handelsbilanz übernommen werden.[111]

Von zentraler Bedeutung für diese Arbeit ist die Behandlung der Managerentlohnung auf Unternehmensebene. Im deutschen Steuerrecht sind die Entlohungszahlungen der Kapitalgesellschaft an den Manager unbeschränkt als Betriebsausgabe abziehbar, zumal sie betrieblich veranlasst sind.[112] Wiederholt kommen in der politischen Diskussion Vorschläge auf, den Abzug von Managergehältern zu begrenzen. Die Entwürfe sehen vor, den vollständigen Abzug bis zu einer Höchstgrenze zuzulassen und darüber hinaus den Abzug ganz oder anteilig zu begrenzen. Typische Höchstgrenzen, die diskutiert werden, liegen bei € 500.000 oder € 1.000.000. Als Vorbild dienen die USA, wo leistungsunabhängige Vergütungsbestandteile über $1.000.000 nicht abzugsfähig sind.[113] Im Analyseteil wird daher sowohl eine unbeschränkte als auch eine beschränkte Abzugsfähigkeit der Vergütungen untersucht.

Durch die bilanzielle Gewinnermittlung stimmt wegen des Periodisierungsprinzips der Zeitpunkt der Entlohnungszahlung nicht immer mit dem Zeitpunkt der Aufwandsentstehung überein. Dies ist für Entlohnungsanteile der Bonusbank von Bedeutung, die der Manager in einer Periode verdient hat, aber nicht ausgezahlt bekommt. Handelsrechtlich kann es sich hierbei um Rückstellungen aus ungewissen Verbindlichkeiten handeln, die passivierungspflichtig sind.[114] Das Handelsrecht ist in diesem Fall maßgeblich für das Steuerrecht, da keine steuerrechtliche Ausnahmeregelung berührt wird.[115]

Für eine Rückstellungsbildung aus ungewissen Verbindlichkeiten müssen mehrere Merkmale erfüllt sein. Zunächst muss es sich um eine Verbindlichkeit aus einer Außenverpflichtung handeln.[116] Dies ist durch die vertragliche Vereinbarung mit dem nicht an der Kapitalgesellschaft beteiligten Manager erfüllt. Außerdem muss eine wirtschaftliche Verursachung der Verpflichtung vor dem Bilanzstichtag gegeben sein.[117] Ähnlich zur Frage, ob beim Manager die begünstigte Fünftelmethode angewandt werden kann, ist eine Abgeltung von Vergangenem für die wirtschaftliche Verursachung zwingend. Dabei reicht ein Anknüpfungspunkt in der Vergangenheit nicht aus.[118]

Hinsichtlich der Zulassung einer Rückstellungsbildung ist der Gesetzgeber strenger als beim Zulassen des begünstigten Steuersatzes beim Manager. Dies wird dadurch deutlich, dass als weiteres Kriterium die "wirtschaftlich wesentlichen Tatbestandsmerkmale [der Verpflichtung] erfüllt" sein müssen und "das Entstehen der Verbindlichkeit nur noch von wirtschaftlich unwesentlichen Tatbestandsmerkmalen"[119] abhängen muss. Davon kann bei einem verzögert ausgezahlten Bonus nicht ausgegangen werden, zumal die Auszahlung von dem Unternehmenserfolg und der Betriebszugehörigkeit des Managers in zukünftigen Perioden abhängt.[120] Die Abgeltung von Vergangenem ist damit nicht gegeben. Daher ist eine Rückstellungsbildung auf Entlohnungsanteile aus Bonusbankvereinbarungen insgesamt nicht zulässig. Alle in der Arbeit betrachtete Vergütungen werden folglich in der Periode der Auszahlung als Betriebsausgabe angesetzt.

Das Wirtschaftsjahr der Kapitalgesellschaft entspricht dem Kalenderjahr. Auf das ermittelte zvE des Wirtschaftsjahres muss die Kapitalgesellschaft einen konstanten Körperschaftsteuersatz von 15% zzgl. Solidaritätszuschlag entrichten.[121] Das nach einkommensteuerlichen Vorschriften ermittelte zvE bildet die Ausgangsgröße zur Ermittlung des Gewerbeertrags.[122] In Prinzipal-Agenten-Anwendungen entsprechen sich der Gewerbeertrag und das zvE typischerweise, weil keine gewerbesteuerlichen Hinzurechnungen (§ 8 GewStG) oder Kürzungen (§ 9 GewStG) vorliegen.[123] Zur Ermittlung der Gewerbesteuerzahllast ist schließlich bei Kapitalgesellschaften der maßgebende Gewerbeertrag mit der einheitlichen Steuermesszahl von 3,5% und dem Gemeindehebesatz zu multiplizieren.[124]

[106] Vgl. FG Nürnberg, Urteil v. 10.05.2017, 3 K 1935/15, Rn. 32.
[107] Siehe § 1 Abs. 1 Nr. 1 KStG und § 2 Abs. 1 GewStG.
[108] Siehe § 1 Abs. 1 SolZG.
[109] Siehe § 6 Abs. 1 HGB i. V. m. § 238 Abs. 1 HGB.
[110] Siehe § 140 AO.
[111] Siehe § 8 Abs. 1 KStG i. V. m. § 5 Abs. 1 Satz 1 EStG.
[112] Siehe § 8 Abs. 1 KStG i. V. m. § 4 Abs. 4 EStG.
[113] Vgl. Voßmerbäumer (2012), S. 537f.
[114] Siehe § 249 Abs. 1 Satz 1 Halbsatz 1 HGB. Vgl. dazu Crezelius (2018), S. 457f, Rn. 117.
[115] Siehe R 5.7 Abs. 1 Satz 1 EStR.
[116] Siehe R 5.7 Abs. 3 EStR.
[117] Siehe R 5.7 Abs. 5 EStR.
[118] Siehe R 5.7 Abs. 5 Satz 3 EStR; dazu Crezelius (2018), S. 461, Rn. 125.
[119] BFH-Urteil v. 01.08.1984, I R 88/80, BStBl. 1985 II, S. 44, Rn. 40.
[120] Vgl. BFH-Urteil v. 18.03.1965, IV 116/64 U, BStBl. III 1965, S. 290, Rn. 10; Zwirner und Tippelhofer (2017/9), S. 10.
[121] Siehe § 23 Abs. 1 KStG.
[122] § 7 Satz 1 GewStG.
[123] Vgl. Jasper (1995), S. 125.
[124] Siehe § 10 Abs. 1 GewStG i. V. m. § 11 Abs. 2 GewStG i. V. m. § 16 GewStG.

Voraussetzung für die Übereinstimmung von zvE und maßgebendem Gewerbeertrag ist, dass die Entlohnungszahlungen an den Manager aus den Unternehmensrückflüssen oder aus vorhandenen liquiden Mitteln finanziert werden und daher keine Fremdfinanzierung notwendig ist. Bei der Ermittlung des zvE sind Zinszahlungen als abziehbare Betriebsausgaben zu behandeln.[125] Allerdings sind ein Viertel der Schuldzinsen aus einer Fremdfinanzierung auf den Gewinn aus Gewerbebetrieb zur Bestimmung der gewerbesteuerlichen Bemessungsgrundlage wieder hinzuzurechnen.[126] Falls Überschüsse oder liquide Mittel im Unternehmen vorliegen und diese am Kapitalmarkt angelegt werden, so stellen die Zinserträge in vollem Umfang steuerpflichtige Einkünfte dar.

Ein weiteres Merkmal des deutschen Steuersystems ist die asymmetrische Behandlung von steuerlichen Gewinnen und Verlusten. Eine Verlustberücksichtigung ist für die Untersuchung von Bonusbanken von Bedeutung, da zu prüfen ist, inwieweit der Prinzipal den Agenten durch die Bonusbank an Verlusten beteiligen kann. Während Gewinne in der Entstehungsperiode unmittelbar zu versteuern sind, so ist eine sofortige Steuererstattung bei steuerlichen Verlusten nur begrenzt vorgesehen. Negative, körperschaftsteuerliche Einkünfte sind lediglich bis zu einem Gesamtbetrag von € 1.000.000 mit positiven Einkünften des vorangegangen Veranlagungszeitraums auszugleichen.[127] Bei gewerbesteuerlichen Verlusten ist hingegen nach geltendem Recht ein Verlustrücktrag nicht möglich.

Nicht rücktragsfähige, negative Einkünfte sind in den folgenden Veranlagungszeiträumen bis zur sogenannten "Mittelstandskomponente" in Höhe von € 1.000.000 unbeschränkt vom Gesamtbetrag der Einkünfte abziehbar. Darüber hinaus ist dies nur für bis zu 60% des € 1.000.000 übersteigenden Betrages möglich.[128] Durch den unvollständigen Verlustvortrag sichert sich der Gesetzgeber eine Mindestbesteuerung von 40% des € 1.000.000 übersteigenden Betrages.[129] Für vorzutragende Gewerbeverluste ist eine der einkommensteuerlichen entsprechende Regelung in § 10a GewStG verankert.[130] Insbesondere sind die Mittelstandskomponente sowie die Mindestbesteuerungsquote gleich hoch.

2.3.3. Auf Anteilseignerebene

Der repräsentative Anteilseigner ist eine natürliche Person, welche in Deutschland unbeschränkt steuerpflichtig gemäß § 1 Abs. 1 EStG ist. Er ist konfessionslos und muss keine Kirchensteuer zahlen. Die 100%-Beteiligung an der Kapitalgesellschaft hält der repräsentative Anteilseigner im Privatvermögen. Die Kapitalgesellschaft dient dem Anteilseigner zur Vertretung seiner Interessen. Besonders sind für ihn die Zahlungsströme relevant, die er aus der Beteiligung am Unternehmen erhält.[131]

Sofern Gewinnausschüttungen der Kapitalgesellschaft getätigt werden, stellen diese als Dividenden Einnahmen aus Kapitalvermögen dar.[132] Unter Vernachlässigung des Sparerpauschbetrages entspricht dies den gesamten Einkünften, die im Rahmen der Partialanalyse zentral sind. Die Einkünfte unterliegen dem konstanten Abgeltungsteuersatz des § 32d Abs. 1 Satz 1 EStG in Höhe von 25% zzgl. Solidaritätszuschlag.[133] Thesaurierte Gewinne, die auf Unternehmensebene verbleiben, unterliegen wegen des steuerlichen Trennungsprinzips keiner Besteuerung auf Anteilseignerebene.

Prinzipiell kann der repräsentative Anteilseigner einen Antrag auf das Teileinkünfteverfahren stellen und die Einkünfte der Veranlagung unterwerfen, weil er zu mehr als 25% an der Kapitalgesellschaft beteiligt ist.[134] Dann würden 60% der Dividendeneinnahmen zum Regeltarif des § 32a EStG veranlagt.[135] Dies wird nachfolgend vernachlässigt, da die 100%-Beteiligung als Hilfskonstrukt dient, um eine klassische Prinzipal-Agenten-Beziehung abzubilden. In Abschnitt 2.1 wurde jedoch deutlich gemacht, dass sich hinter dem repräsentativen Anteilseigner viele Investoren mit homogenem Interesse verbergen. Weiterhin kann dies damit begründet werden, dass der relevante Grenzeinkommensteuersatz bei dem Prinzipal so hoch ist, dass eine Antragssetzung unvorteilhaft ist.[136]

Zu welchem Zeitpunkt die steuerliche Belastung beim Anteilseigner stattfindet, hängt von der Ausschüttungspolitik der Kapitalgesellschaft ab.[137] Grundsätzlich ist zu beachten, dass handelsbilanzielle Ausschüttungssperrvorschriften existieren, die eine Ausschüttung der im Unternehmen vorhandenen liquiden Mittel verhindern können.[138] Darüber hinaus ist im Aktiengesetz eine spezielle Ausschüttungssperre vorgesehen, sofern gesetzliche Rücklagen und Kapitalrücklagen zusammen nicht 10% des Grundkapitals übersteigen.[139] Diesbezüglich wird mit Blick auf das Partialkalkül festgestellt, dass stets von einem ausreichend hohen Ausschüttungspotenzial ausgegangen wird.

2.4. Grundstruktur einer Prinzipal-Agenten-Beziehung

Die Beziehung zwischen dem Manager eines Unternehmens und dem Anteilseigner wurde bereits als typisches Beispiel einer Prinzipal-Agenten-Beziehung identifiziert. Ausgangspunkt für diese ist der Vertrag zwischen den beiden

[125] Siehe § 8 Abs. 1 Satz 1 KStG i. V. m. § 4 Abs. 4 EStG. Von der Betrachtung der Zinsschranke (§ 8a KStG) wird abgesehen.
[126] Siehe § 8 Nr. 1 Buchstabe a GewStG. Der Freibetrag in Höhe von € 100.000 gilt im Sinne des Partialkalküls als ausgeschöpft.
[127] Siehe § 8 Abs. 1 KStG i. V. m. § 10d Abs. 1 Satz 1 EStG.
[128] Siehe § 8 Abs. 1 KStG i. V. m. § 10d Abs. 2 Satz 1 EStG.
[129] Vgl. Niemann (2004), S. 362.
[130] Siehe § 10a Sätze 1 und 2 GewStG.
[131] Vgl. König und Wosnitza (2004), S. 78.
[132] Siehe § 20 Abs. 1 Nr. 1 EStG.
[133] Der Zuschlagssatz beträgt hier ebenfalls 5,5% der festzusetzenden Einkommensteuer.
[134] Siehe § 32d Abs. 2 Nr. 3 Buchstabe a EStG.
[135] Siehe § 3 Nr. 40 Buchstabe d EStG.
[136] Unter Vernachlässigung von Werbungskosten aus der Beteiligung wäre dies der Fall, wenn der Anteilseigner einem Grenzsteuersatz von 42% unterliegt, da $0,6 \times 0,42 = 0,252 > 0,25$.
[137] Siehe dazu die Prämisse in Kapitel 4.1.2.
[138] Vgl. König und Wosnitza (2004), S. 87.
[139] Siehe § 150 Abs. 3 AktG.

Akteuren, der die Rechte und Pflichten abhängig von sämtlichen innerhalb der Vertragsdauer auftretenden Eventualitäten regelt. Der Agent verpflichtet sich mit Annahme des Vertrages, für den Prinzipal zu arbeiten. Außerdem spezifiziert der Vertrag die Aufteilung des Unternehmenserfolges, typischerweise des Gewinns, zwischen den Akteuren. Während der Agent eine Kompensation erhält, die abhängig vom Eintreten vertraglich vereinbarter Zustände ist, bekommt der Prinzipal die Residualgröße aus Unternehmenserfolg und Kompensation. Entscheidend für die Durchsetzbarkeit des Vertrages ist die Verifizierbarkeit der jeweiligen Zustände. Dies bedeutet, dass eine dritte Partei wie ein Gericht die Vertragsbedingungen nach Vertragsabschluss durchsetzen kann und ein Vertragsbruch damit nicht möglich ist. Die Verifizierbarkeit ist deshalb von Bedeutung, weil beide Parteien ihren individuellen Nutzen maximieren und es dadurch zum Interessenkonflikt kommt.[140]

Im Modellteil wird der Konflikt zwischen Manager und Anteilseigner lediglich auf die Höhe des Anstrengungsniveaus des Agenten reduziert. Der Prinzipal möchte einen maximalen, nachsteuerlichen Kooperationsgewinn erzielen. Der Agent muss hierfür eine hohe Anstrengung leisten, da er durch seine Tätigkeit den Cash-Flow des Unternehmens beeinflusst.[141] Gleichzeitig strebt der Prinzipal eine möglichst geringe Kompensation des Agenten an, um dieses hohe Aktivitätsniveau zu induzieren. Der Manager will auf der anderen Seite eine hohe Nettokompensation erzielen. Gleichzeitig möchte er den damit verbundenen Arbeitseinsatz gering halten, zumal dieser für ihn Anstrengungskosten verursacht. Der Prinzipal muss nun bei der Vertragsgestaltung das strategische Verhalten des Agenten antizipieren und die Wechselwirkungen zwischen einer höheren Kompensation und dem damit verbundenen marginalen Effekt auf den Arbeitseinsatz und den Unternehmenserfolg berücksichtigen.[142]

Die zeitliche Abfolge der Interaktion zwischen Prinzipal und Agent im Grundfall wird in Abbildung 4 dargestellt.
Zu Beginn erfolgt das Vertragsangebot durch den Prinzipal. Der Vertrag wird im Grundmodell nicht nachverhandelt und der Agent kann sich nur dazu entscheiden, diesen anzunehmen oder abzulehnen. Der Agent nimmt den Vertrag an, wenn er einen Mindestnutzen aus den zu erwartenden Vergütungshöhen erhält, den er ebenfalls aus einer Alternativbeschäftigung erhalten würde. Dieser Mindestnutzen wird als *Reservationsnutzen*, die korrespondierende Bedingung im Optimierungsproblem des Prinzipals als *Partizipations-* oder *Teilnahmebedingung* bezeichnet. Außerdem muss das Vergütungssystem derart gestrickt sein, dass der Agent die vom Prinzipal gewünschte bzw. antizipierte Arbeitsanstrengung wählt. Dies wird im Optimierungsprogramm des Prinzipals in der *Anreizkompatibilitätsbedingung* sichergestellt.[143] Hierbei kann der Prinzipal den Agenten zu jeder beliebigen Anstrengungshöhe motivieren, muss den Vorteil der höheren Anstrengung allerdings gegen die erhöhten Kompensationskosten abwägen.[144]

Von großer Bedeutung ist bei der Vertragsgestaltung der Informationsstand der beiden Akteure. Im Rahmen dieser Arbeit wird davon ausgegangen, dass beide Akteure vor Vertragsabschluss homogene Informationen haben. Zudem erhält der Agent nach Vertragsabschluss keine privaten Informationen über den Umweltzustand.[145] Als wesentliches Anreizproblem dieser Arbeit verbleibt damit die asymmetrische Informationsverteilung zwischen Prinzipal und Agent nach Vertragsabschluss, bei der der Prinzipal die Handlung des Agenten nicht beobachten kann *(hidden action)*. Hierdurch entsteht das Problem des moralischen Risikos.[146] Diese Art der Informationsasymmetrie hat in der wissenschaftlichen Diskussion zu steuerlichen Auswirkungen die größte Aufmerksamkeit erfahren.

Nimmt der Agent den Vertrag an, so wählt er im Anschluss eine Arbeitsanstrengung für die gesamte Periode.[147] Homogene Informationen vor Vertragsabschluss implizieren, dass beide Parteien dieselben Informationen bezüglich des Zusammenhangs von Arbeitsanstrengung, Cash-Flow und exogenen Umwelteinflüssen haben. Ist das Aktivitätsniveau des Agenten beobachtbar und verifizierbar, so besteht kein Anreizkonflikt, da der Prinzipal den Vertrag bedingt auf die von ihm gewünschte Anstrengungswahl formulieren kann. In diesem Rahmen kommt es zu einem First-best-Vertrag, der aus Prinzipalsicht optimal ist. Die Unbeobachtbarkeit der Agentenhandlung wäre einer First-best-Situation nicht hinderlich, wenn der Prinzipal den exogenen Umweltfaktor, der darüber hinaus auf den Output wirkt, beobachten könnte. Dann könnte er den am Ende der Periode realisierten Cash-Flow betrachten und auf die Aktion des Agenten zurückschließen. Folglich entstehen erst durch den ebenfalls unbeobachtbaren Umweltzustand Agenturkosten. Dieses klassische Rückschlussproblem ist eine notwendige Bedingung dafür, dass nur eine Second-best-Lösung möglich ist.[148]

Zum Ende der Periode realisiert sich der durch die Agentenhandlung beeinflusste Cash-Flow und mithin die Entlohnung des Agenten. Diese bestimmen die Steuerbemessungsgrundlagen auf den unterschiedlichen Ebenen eindeutig und führen zu den Steuerzahlungen. Die Nettoentlohnung des Agenten bzw. die Nettodividende des Prinzipals steht nun der eigentlichen Zielsetzung der beiden Akteure, dem aus den überschüssigen Zahlungen entstehenden Konsum, zur Verfügung.[149]

[140] Vgl. Jost (2001a), S. 13ff.

[141] Die Begriffe Cash-Flow und Output werden im Folgenden als Synonyme genutzt.

[142] Vgl. Jost (2001a), S. 16f.

[143] Abkürzend wird diese in der Arbeit auch *Anreizbedingung* genannt.

[144] Vgl. Jost (2001a), S. 18f.

[145] Damit werden *hidden characteristics-* und *hidden information-*Situationen ausgeschlossen.

[146] Vgl. Schäfer (2013), S. 10.

[147] Von mehreren Aufgaben des Agenten wird in der Arbeit abgesehen. In einem Multi-Task-Fall müsste der Agent seine Anstrengung auf mehrere Aufgaben verteilen. In der Steuerliteratur wurde neben den operativen Tätigkeiten des Agenten auch die simultane Durchführung von Steuervermeidungspraktiken diskutiert. Vgl. Ewert und Niemann (2012), S. 22ff.

[148] Vgl. Laffont und Martimort (2009), S. 149f; Jost (2001a), S. 21.

[149] Vgl. Laux (2003), S. 37.

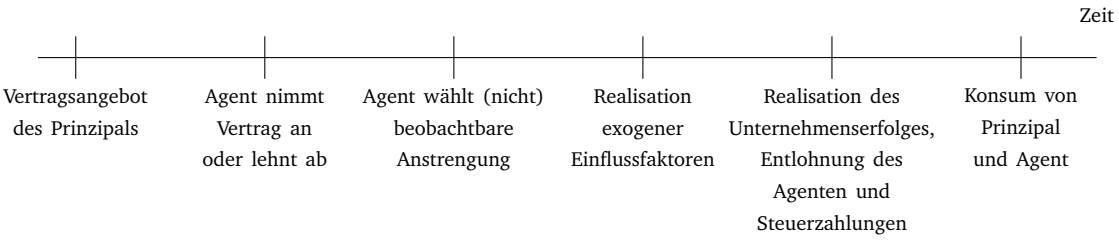


Abbildung 4: Zeitliche Abfolge einer klassischen Prinzipal-Agenten-Beziehung (In Anlehnung an Jost, 2001a, S. 27)

Letztlich ist mit den Risikopräferenzen der Akteure ein weiterer Modellierungsaspekt von großer Bedeutung für die Prinzipal-Agenten-Beziehung, da dieser die Gründe für Effizienzverluste im Rahmen eines Second-best-Vertrages determiniert. Unstrittig ist die Unterstellung eines risikoneutralen Prinzipals hinsichtlich seines Nachsteuergewinns. Dahinter steckt die Überlegung, dass Unternehmenseigner durch Beteiligungen an verschiedenen Unternehmen ihr Unternehmensrisiko diversifizieren und damit die Konsummöglichkeiten glätten können, wobei dies beim Manager nicht der Fall ist.[150] Deshalb ist in den Pionierarbeiten zu moralischem Risiko stets von einem risikoaversen Agenten ausgegangen worden.[151]

Unter Berücksichtigung von Steuern wurde häufig das LEN-Modell mit risikoaversem Agenten nach Spremann (1987) zur Analyse genutzt.[152] Neben einigen Vorteilen ist die Nutzung des LEN-Modells sowie die Annahme eines risikoaversen Agenten in dieser Arbeit nicht praktikabel. Zum einen liegt dies daran, dass die Einbeziehung der steuerlich asymmetrischen Behandlung von Gewinnen und Verlusten auf Unternehmensebene die Linearitätsannahme im LEN-Modell verletzt.[153] Zum anderen besteht die grundsätzliche Schwierigkeit, allgemeine Eigenschaften aus mehrperiodigen Prinzipal-Agenten-Modellen mit risikoaversem Agenten zu ziehen. Eine analytische Lösbarkeit ist nur für konkrete Nutzenfunktionen möglich, wodurch die Allgemeingültigkeit der Ergebnisse verloren geht. Außerdem ist in Frage zu stellen, ob Vorstandsvorsitzende von (großen) Kapitalgesellschaften tatsächlich risikoavers sind. Insbesondere in der juristischen Literatur werden diese als "unternehmerähnliche Personen, die [...] keiner Fürsorge oder Risikoentlastung bedürfen"[154], bezeichnet.

Deshalb erscheint es sinnvoll, im Folgenden Risikoneutralität des Agenten hinsichtlich seiner Nettoentlohnung zu unterstellen. Risikoteilungsaspekte zwischen Prinzipal und Agent spielen somit keine Rolle. Um eine triviale Lösung bei Risikoneutralität zu vermeiden, ist es darüber hinaus notwendig, eine Haftungsbeschränkung des Agenten vorauszusetzen. Eine Haftungsbeschränkung schützt den Agenten vor Sanktionszahlungen bei schlechten Unternehmensergebnissen, die er jedoch ggf. leisten könnte.[155] Die Haftungsbeschränkung ist insgesamt eine geeignete Annahme, da in Kapitel 2.2 ohnehin verdeutlicht wurde, dass Rückzahlungsverpflichtungen in der Praxis arbeitsrechtlich nicht möglich sind und eine mangelnde Verlustbeteiligung des Managers zum Interessenkonflikt der beiden Parteien führt. Der Prinzipal könnte ohne die Haftungsbeschränkung mit einer Entlohnung in Abhängigkeit vom Cash-Flow die First-best-Lösung erzielen, indem er den Agenten für einen hohen Rückfluss belohnt und für einen niedrigen sanktioniert. Durch den Ausschluss der Sanktion wird der Handlungsspielraum des Prinzipals auf positive Bonuszahlungen beschränkt, wodurch er für eine hinreichende Anreizsetzung eine Rente an den Agenten zahlen muss.[156]

3. Steuerwirkungen in einperiodigen Prinzipal-Agenten-Beziehungen

In diesem Kapitel wird mit der Darstellung des einperiodigen Prinzipal-Agenten-Modells die Grundlage für die Analyse von steuerlichen Einflüssen auf langfristige Bonusbankvereinbarungen in Kapitel 4 gelegt. Hierzu wird das in der Arbeit genutzte Modell formalisiert. In der formalen Modellierung wird auf die explizite Darstellung des Grundmodells ohne Steuern verzichtet, da das gegebene Steuersystem die ökonomische Realität der Akteure beschreibt. Die Analyse der Auswirkungen von Steuern aus Sicht des Prinzipals werden bei beobachtbarer und unbeobachtbarer Anstrengung des Agenten in den Mittelpunkt gestellt.

3.1. Modellierungswahl und steuerliche Prämissen

Das einperiodige Modell kombiniert wesentliche Aspekte aus der Prinzipal-Agenten-Literatur zu steuerlichen Anreizwirkungen und schafft hierdurch einen einheitlichen Rahmen zur Analyse. Insbesondere fließen Modellierungsaspekte der Untersuchungen von Meißner, Schneider und Scholze (2014) und Niemann (2011) in das Modell mit

[150] Vgl. Niemann und Simons (2003), S. 24, Fußnote 2.
[151] Vgl. z. B. Holmström (1979), S. 75f.
[152] Vgl. Bauer et al. (2018) S. 37. LEN steht für eine lineare Vergütung, exponentielle Nutzenfunktion und normalverteilte Umwelteinflüsse.
[153] Vgl. Niemann (2008), S. 288f. Der Einfluss von Verlusten ist im Kapitel 4.4 von Interesse.
[154] Martens (2005), S. 127.
[155] Vgl. Gillenkirch (1997), S. 8.
[156] Vgl. Laffont und Martimort (2009), S. 149.

ein. Beide Aufsätze untersuchen steuerinduzierte Effekte im Hinblick auf die Anreizsetzung des Prinzipals. Meißner et al. vergleichen dabei den Einfluss einer Bonussteuer auf Unternehmens- und Managerebene, während Niemann die asymmetrische Behandlung von steuerlichen Gewinnen und Verlusten analysiert. Beiden Aufsätzen ist gemein, dass sie Modelle mit binärem Cash-Flow und binärem Anstrengungsniveau nutzen. Dieser Modellrahmen wird nachfolgend ebenfalls genutzt.

Im Modell dieser Arbeit ist der Prinzipal zu 100% an der eigenen Kapitalgesellschaft beteiligt und beauftragt den Agenten über eine Periode mit der Unternehmensleitung. Der Agent kann ein hohes oder ein niedriges Anstrengungsniveau $a \in \{a^L, a^H\}$ wählen und damit den stochastischen Cash-Flow vor Steuerzahlung des Unternehmens $\tilde{x} \in \{x^L, x^H\}$ mit $x^H > x^L \geq 0$ beeinflussen. Die Realisation von $\tilde{x}$ ist stets beobachtbar und kontrahierbar. Sowohl dem Prinzipal als auch dem Agenten ist der Einfluss der Anstrengung auf den Output bekannt. Die Wahrscheinlichkeit, einen hohen Output zu erzielen, wenn der Agent ein hohes Aktivitätsniveau wählt, beträgt $Pr(\tilde{x} = x^H | a = a^H) = \phi^H$. Dagegen beträgt die entsprechende Wahrscheinlichkeit im Falle eines niedrigen Aktivitätsniveaus $Pr(\tilde{x} = x^H | a = a^L) = \phi^L$. Sinnvollerweise ist $\phi^H > \phi^L$ erfüllt, sodass bei einem hohen Rückfluss ein höheres Anstrengungsniveau eher zu erwarten ist als ein niedriges. Diese Annahme ist in der Literatur als *Monotone Likelihood Ratio Property* bekannt.[157] Zudem gilt $\phi^j \in (0,1)$, wobei $j \in \{H, L\}$ der Index für das hohe bzw. niedrige Anstrengungsniveau ist. Der Prinzipal kann damit nie vom realisierten Cash-Flow auf das Aktivitätsniveau des Agenten schließen, da mit Wahrscheinlichkeit $1 - \phi^H > 0$ (bzw. $1 - \phi^L > 0$) eine hohe (bzw. niedrige) Anstrengung zum niedrigen Rückfluss führt. Damit ist jede Outputhöhe bei jedem Anstrengungsniveau möglich.

Das Aktivitätsniveau ist ausschließlich für den Agenten selbst beobachtbar. Der Prinzipal hat in einem Modell mit binärer Anstrengung drei Möglichkeiten:[158] Er kann die Kooperation mit dem Agenten nicht eingehen (Unterlassungsalternative), ihm lediglich einen Fixlohnvertrag mit einem Grundgehalt F anbieten, der den Agenten zur Wahl von Anstrengung a^L bewegt oder eine Bruttoentlohnung $w(\tilde{x}) = \tilde{w} \in \{w^L, w^H\}$ in Abhängigkeit vom realisierten Cash-Flow wählen, um Anstrengung a^H zu induzieren. Das Optimierungsprogramm des Prinzipals kann hierbei in zwei Schritte zerlegt werden. Zunächst bestimmt er für jedes Anstrengungsniveau den kostengünstigsten Vertrag, der dieses Anstrengungsniveau implementiert. Von allen Alternative wählt er diejenige mit dem größten Zielfunktionswert.[159]

Der risikoneutrale Prinzipal besitzt kein privates Vermögen und maximiert sein Endvermögen nach Steuern EV aus der Kooperation. Alle Zahlungen werden folglich auf das Ende des Kooperationszeitraums bezogen.[160] Die Kompensation des Agenten erfolgt zum Ende der ersten Periode genauso wie die daraufhin folgende Ausschüttung der Überschüsse aus dem Unternehmen an den Prinzipal. Die konkreten Entlohnungen und Endvermögenswerte für den nachsteuerlichen Fall werden für die Alternativen an späterer Stelle bestimmt. Die Unterlassungsalternative liefert dem Prinzipal jedoch einen Endvermögenswert von $EV = 0$, der stets durch die Endvermögenshöhen bei Kooperation mit dem Agenten dominiert wird.[161] Eine Rangfolge der Endvermögenshöhen bei Implementierung von a^H bzw. a^L kann hingegen nicht allgemein bestimmt werden und muss daher parameterabhängig geprüft werden.

Bei hoher Anstrengung erleidet der Agent einen Disnutzen von v^H, bei niedriger Anstrengung einen Disnutzen von v^L mit $v^H > v^L \geq 0$. Diese sind als monetäres Nutzenäquivalent gegeben. Da keine Zahlungswirkungen aus den Disnutzen resultieren, sind sie unabhängig von steuerlichen Einflussgrößen. Der Agent ist risikoneutral und unterliegt einer beschränkten Haftung mit Haftungsvermögen $W = 0$. Dies impliziert, dass ein Malus nicht möglich ist und die Bruttoentlohnungen w^H, w^L und $F \geq 0$ sein müssen. Darüber hinaus ist der Agent beschränkt benevolent und wählt damit die hohe Anstrengung, sofern er indifferent zwischen einer hohen und niedrigen Anstrengung ist.

Beim Vertragsangebot muss der Prinzipal beachten, dass der Agent einen Reservationsnutzen $\underline{U} \geq 0$ fordert, der die beste, alternative Beschäftigungsmöglichkeit des Agenten widerspiegelt. Dabei muss grundsätzlich davon ausgegangen werden, dass der Reservationsnutzen abhängig von steuerlichen Parametern auf Managerebene ist, da die Einkünfte aus der Alternativtätigkeit ebenfalls steuerbar sind.[162] Die Art der Besteuerung der Alternative kann aber in dieser Arbeit nicht angegeben werden, weil die Alternative nicht explizit modelliert wird.[163] Hierzu wären einige Annahmen hinsichtlich des mit der Alternativbeschäftigung verbundenen Arbeitsleids sowie der Vergütungsart und der damit verbundenen Steuerlast erforderlich.[164] Im Rahmen des nachfolgenden Modells ist dies nicht abbildbar und es wird von einem Reservationsnutzen unabhängig von der Besteuerung ausgegangen.

Im Sinne einer multilateralen Analyse werden Ertragsteuern auf Ebene der Kapitalgesellschaft, des Prinzipals und des Agenten integriert.[165] Es erfolgt eine Grenzbetrachtung, so-

[157] Vgl. Milgrom (1981), S. 382f.
[158] Vgl. Niemann (2011), S. 5.
[159] Vgl. Grossman und Hart (1983), S. 12f.
[160] Vgl. Blohm, Lüder und Schäfer (2012), S. 72ff.
[161] Ein Endvermögen von Null liegt vor, wenn auf Unternehmensebene keine liquiden Mittel L vorliegen oder diese wegen ihrer Entscheidungsirrelevanz vernachlässigt werden können. Letzteres ist hier der Fall. Zur weiteren Begründung sei auf das Ende dieses Abschnittes verwiesen.
[162] Eine Ausnahme liegt nur vor, wenn der Reservationsnutzen das Nutzenniveau des Existenzminimums widerspiegelt, welches in Deutschland steuerfrei ist oder durch Transferleistungen garantiert ist. Vgl. Ewert und Niemann (2012), S. 10. Dass die Alternative einer Führungskraft dieses Existenzminimum ist, erscheint jedoch zweifelhaft.
[163] Vgl. Krenn (2017), S. 509ff zur expliziten Modellierung der Alternativbeschäftigung.
[164] Vgl. Ewert und Niemann (2012), S. 10.
[165] Der multilaterale Planungsansatz geht nach Bauer et al. auf Scholes, Wolfson, Erickson, Maydew und Shevlin (2009) zurück. Vgl. Bauer et al. (2018), S. 35.

dass persönliche Umstände wie Freibeträge und -grenzen, Sonderausgaben oder Pauschbeträge bei allen Akteuren nicht in die Analyse miteinbezogen werden. Auf Ebene der Kapitalgesellschaft stellt der unsichere Cash-Flow $\tilde{x}$ eine laufende Einnahme aus dem Geschäftsbetrieb dar und unterliegt neben dem effektiven Körperschaftsteuersatz s^{KSt}, der den Solidaritätszuschlag beinhaltet, auch dem Gewerbesteuersatz s^{GewSt}, der sich durch Multiplikation der Steuermesszahl mit dem Gemeindehebesatz ergibt.

Im deutschen Steuerrecht stellen die Bruttoentlohnungen w^H, w^L und F auf Unternehmensebene abzugsfähige Betriebsausgaben dar. Da eine beschränkte Abzugsfähigkeit von Managervergütungen in der politischen Diskussion wiederholt aufgekommen ist, wird wie in Bauer et al. (2018) ein nicht von der Unternehmensbemessungsgrundlage abzugsfähiger Teil der Bruttoentlohnung im Parameter $k \in [0,1)$ erfasst. Dieser gilt sowohl für die Körperschaftsteuer als auch für die Gewerbesteuer, wobei für $k = 0$ die deutsche Rechtslage abgebildet wird.[166] Außerdem dürfen sämtliche Steuerzahlungen das zvE bzw. den Gewerbeertrag nicht mindern, sodass als Bemessungsgrundlage $\tilde{x} - (1-k)\tilde{w}$ bzw. $\tilde{x} - (1-k)F$ verbleibt. Zwischen dem zvE und dem Gewerbeertrag besteht kein Unterschied, zumal durch die Vertragsbeziehung keine gewerbesteuerlichen Hinzurechnungs- oder Kürzungsvorschriften berührt werden. Insbesondere erfolgt keine Fremdfinanzierung der Entlohnungen. Daher kann davon ausgegangen werden, dass die Kapitalgesellschaft einen kombinierten Ertragsteuersatz in Höhe von $\tau_c \in [0,1)$ mit $\tau_c = s^{KSt} + s^{GewSt}$ abzuführen hat.

Zum Ende der Periode schüttet die Kapitalgesellschaft sämtliche in der Periode erwirtschafteten Überschüsse aus. Auf Ebene des Prinzipals unterliegen diese der proportionalen Einkommensteuer zuzüglich Solidaritätszuschlag auf Dividenden $\tau_d \in [0,1)$. Das Endvermögen des Prinzipals nach Steuern lautet damit bei hoher induzierter Anstrengung:

$$\begin{aligned} EV\left(a^H\right) &= (1-\tau_d)\big[(1-\tau_c) \\ &\quad \left(\phi^H\left(x^H - w^H\right) + \left(1-\phi^H\right)\left(x^L - w^L\right)\right) \\ &\quad - k\tau_c\left(\phi^H w^H + \left(1-\phi^H\right)w^L\right)\big] \\ &= (1-\tau_d)\left[(1-\tau_c)E(\tilde{x}-\tilde{w}) - k\tau_c E(\tilde{w})\right] \end{aligned} \tag{1}$$

E ist der Erwartungswertoperator und $k\tau_c E(\tilde{w})$ bildet die zusätzlich erwartete Steuerlast aufgrund der beschränkten Abzugsfähigkeit der Managervergütung ab. Analog gilt für das Endvermögen bei Implementierung von a^L und der Entlohnung F:

$$\begin{aligned} EV\left(a^L\right) &= (1-\tau_d) \\ &\quad \left[(1-\tau_c)\left(\phi^L x^H + \left(1-\phi^L\right)x^L - F\right) - k\tau_c F\right] \\ &= (1-\tau_d)\left[(1-\tau_c)E(\tilde{x}-F) - k\tau_c F\right] \end{aligned} \tag{2}$$

Hierbei ist zu beachten, dass die im Erwartungswert zugrunde liegenden Wahrscheinlichkeiten abhängig von der induzierten Anstrengung sind. Der ausschüttungsfähige Betrag sei unabhängig von dem realisierten Output nicht negativ. Dies erfordert $x^j \geq (1+\frac{k\tau_c}{1-\tau_c})w^j$ für $j \in \{L,H\}$ und $x^L \geq (1+\frac{k\tau_c}{1-\tau_c})F$.[167] Liquide Mittel $L \geq 0$ auf Unternehmensebene sind damit nicht entscheidungsrelevant, da diese nicht zur Entlohnungsfinanzierung benötigt werden. Steuerliche Verluste auf Unternehmensebene sind durch den positiven, ausschüttungsfähigen Betrag ausgeschlossen.[168]

Die Entlohnungen w^H, w^L und F unterliegen auf Managerebene der tariflichen Einkommensteuer zuzüglich Solidaritätszuschlag. Sie wird durch einen proportionalen Einkommensteuersatz $\tau_{inc} \in [0,1)$ approximiert. Dies ist allenfalls unkritisch, sofern das zvE des Managers unabhängig von den Entlohnungen durch andere Einkünfte stets in der selben Proportionalzone des § 32a EStG landet. Diese vereinfachende Annahme wird getroffen, da ansonsten der relevante Grenzsteuersatz vom zvE bzw. den Entlohnungen abhängen würde. Dies erschwerte die Analyse wesentlich.[169] Als Ausgangspunkt der Analyse wird im nächsten Abschnitt das Optimierungsproblem des Prinzipals aufgezeigt.

3.2. Optimierungsproblem und Entlohnungswahl des Prinzipals bei unbeobachtbarer Arbeitsanstrengung

Für die Analyse ist die Annahme sinnvoll, dass vor Steuern a^H effizient ist (Ex ante Effizienzbedingung), d. h. es gilt

$$\begin{aligned} &\phi^H x^H + \left(1-\phi^H\right)x^L - v^H > \\ &\phi^L x^H + \left(1-\phi^L\right)x^L - v^L \\ \Leftrightarrow &\underbrace{\left(x^H - x^L\right)}_{:=\Delta x}\underbrace{\left(\phi^H - \phi^L\right)}_{:=\Delta\phi} > \underbrace{v^H - v^L}_{:=\Delta v}. \end{aligned} \tag{3}$$

Wie jedoch nachfolgend deutlich wird, führt dies nicht zwangsläufig dazu, dass der Prinzipal die hohe Anstrengung induzieren möchte. Das Optimierungsproblem des Prinzipals, bei dem er die hohe Anstrengung des Agenten motiviert, lautet:

$$\max_{w^L, w^H} EV\left(a^H\right)$$

$$\text{s.t. } (1-\tau_{inc})\left(\phi^H w^H + \left(1-\phi^H\right)w^L\right) - v^H \geq \underline{U}, \tag{4}$$

$$(1-\tau_{inc})\left(\phi^H w^H + \left(1-\phi^H\right)w^L\right) - v^H \geq (1-\tau_{inc})\left(\phi^L w^H + \left(1-\phi^L\right)w^L\right) - v^L, \tag{5}$$

$$w^L, w^H \geq 0 = W. \tag{6}$$

Der Prinzipal wählt die Vergütungshöhen so, dass sein Endvermögen maximal wird. Dabei stellt er die Partizipation des

[166] Vgl. Bauer et al. (2018), S. 40f.

[167] Die Beschränkungen ergeben sich durch Umformen der realisierten Bemessungsgrundlagen der Dividendensteuer aus (1) und (2). Bspw. $(1-\tau_c)(x^L - F) - k\tau_c F \geq 0 \Leftrightarrow x^L \geq (1+\frac{k\tau_c}{1-\tau_c})F$.

[168] Dies folgt wegen $x^j \geq (1+\frac{k\tau_c}{1-\tau_c})w^j \geq (1-k)w^j$ für $j \in \{L,H\}$ und $x^L \geq (1+\frac{k\tau_c}{1-\tau_c})F \geq (1-k)F$.

[169] Die Wirkung eines direkt progressiven Steuersatzes ist ähnlich zu einem Agenten, der risikoavers bezüglich seines Bruttoeinkommens in einer Welt ohne Steuern ist. Vgl. Jasper (1995), S. 68. Risikoaversion wurde bereits zur Komplexitätsreduktion ausgeschlossen.

Agenten durch die erste Nebenbedingung (4) sicher, in dem die nachsteuerliche Entlohnung bei hoher Anstrengung im Erwartungswert mindestens dem Reservationsnutzen und dem Arbeitsleid entspricht. Die zweite Nebenbedingung (5) garantiert, dass der Agent zum gegebenen Entlohnungssystem $\left(w^L, w^H\right)$ nicht die niedrige Anstrengung a^L wählt (Anreizkompatibilität). Die Haftungsbeschränkung in Ungleichung (6) gewährleistet, dass die Entlohnungen nicht negativ werden.

Eine Umformung der Anreizkompatibilitätsbedingung (5) führt zu

$$\Delta w = w^H - w^L \geq \frac{v^H - v^L}{\phi^H - \phi^L} \frac{1}{1-\tau_{inc}} . \tag{7}$$

Die Bedingung (7) verdeutlicht, dass zur Sicherung der Anreizkompatibilität lediglich die Entlohnungsspanne Δw von Bedeutung ist. Die Interpretation ist intuitiv: Je kleiner (bzw. größer) die Differenz der Disnutzen (bzw. der bedingten Wahrscheinlichkeiten für einen hohen Output) ist, desto kleiner kann die Entlohnungsdifferenz sein, um eine hohe Anstrengung des Agenten zu motivieren. Der Prinzipal wählt hierbei im Optimum die Entlohnung w^L so klein wie möglich, da diese keine Auswirkung auf das Anstrengungsniveau des Agenten hat. Aufgrund der Haftungsbeschränkungen in der Nebenbedingung (6) setzt er daher die Entlohnung bei niedrigem Cash-Flow auf $w^L = 0$.[170] Weil der Agent beschränkt benevolent ist, wählt der Prinzipal die Entlohnung w^H so, dass die Nebenbedingung (7) mit Gleichheit erfüllt ist. Daraus folgt unmittelbar $w^H = \frac{v^H - v^L}{\phi^H - \phi^L} \cdot \frac{1}{1-\tau_{inc}}$. Damit entspricht in der Second-best-Lösung aufgrund des Haftungsvermögens von $W = 0$ die Entlohnung w^H der Entlohnungsspanne Δw.

Die Haftungsbeschränkung sowie die unbeobachtbare Anstrengung des Agenten führen dazu, dass die in Kapitel 2.4 dargestellten First-best-Lösungen nicht erreicht werden.[171] Allerdings erweist es sich in diesem Modellrahmen bei unbeobachtbarer Arbeitsanstrengung darüber hinaus als notwendig, weitere Größeneinschränkungen bei den Parametern v^L und $\underline{U}$ vorzunehmen, da die First-best-Lösung ansonsten trotz unbeobachtbarer Handlung und Haftungsbeschränkung eintreten kann.[172] Um diesen irrelevanten Fall auszuschließen, der letztlich mit einem fehlenden Konflikt zwischen Prinzipal und Agent gleichzusetzen ist, wird die Teilnahmebedingung des Agenten (4) nun detaillierter betrachtet. Eine Umformung dieser unter Berücksichtigung der optimalen Entlohnungshöhen aus Prinzipalsicht von $w^L = 0$ und $w^H = \frac{v^H - v^L}{\phi^H - \phi^L} \cdot \frac{1}{1-\tau_{inc}}$ liefert

$$(1-\tau_{inc})\,\phi^H w^H = \phi^H \frac{v^H - v^L}{\phi^H - \phi^L} \gtreqless \underline{U} + v^H. \tag{8}$$

Es kann nicht eindeutig festgelegt werden, ob die linke oder die rechte Seite in der Bedingung (8) größer ist. Dies ist abhängig von den konkreten Parameterausprägungen. Im Fall $\phi^H \frac{v^H - v^L}{\phi^H - \phi^L} < \underline{U} + v^H$ müsste der Prinzipal dem Agenten ein zusätzliches Grundgehalt $w^L > 0$ zahlen, damit die Partizipationsbedingung überhaupt erfüllt ist. Dieses Grundgehalt ist notwendig, wenn $\underline{U}$ bzw. v^L sehr hoch sind. Im ersten Fall liegt dem zugrunde, dass $\underline{U}$ relativ zur erwarteten Bonuszahlung hoch ist, wodurch das Grundgehalt zur Teilnahme des Agenten benötigt wird. Im zweiten Fall macht ein höheres Arbeitsleid bei niedriger Anstrengung v^L es dem Prinzipal leichter, ein hohes Anstrengungsniveau beim Agenten zu motivieren, da die Disnutzendifferenz Δv kleiner wird. Deshalb ist die resultierende Bonuszahlung w^H im Optimum relativ zu $\underline{U} + v^H$ klein, sodass ebenfalls das Grundgehalt zur Vertragsannahme des Agenten notwendig ist. Wenn dieses strikt positive Grundgehalt bei Implementierung von a^H benötigt wird, führt dies unmittelbar zur First-best-Lösung. Sie ist durch bindende Nebenbedingungen (4) und (5) und eine nicht bindende Haftungsbeschränkung (6) charakterisiert.

Eine hinreichende Bedingung zum Ausschluss dieses irrelevanten Falles ist $\underline{U} = v^L = 0$. Hieraus folgt nach Bedingung (8) unmittelbar $\frac{\phi^L v^H}{\phi^H - \phi^L} > 0$. Für die nachfolgenden Überlegungen werden $\underline{U}$ und v^L allgemein gehalten. Allerdings wird

$$\phi^H \frac{v^H - v^L}{\phi^H - \phi^L} > \underline{U} + v^H \tag{9}$$

unterstellt, sodass die nötige Entlohnungsspanne zur Sicherstellung der hohen Anstrengung so groß ist, dass der Agent eine Rente erzielt. Diese Rente gleicht den Agenturkosten aus der asymmetrischen Informationsverteilung.

Bei Implementierung der Anstrengung a^L ist diese Bedingung nicht von Bedeutung. Das Optimierungsproblem des Prinzipals lautet

$$\max_F EV\left(a^L\right)$$

$$\text{s.t. } (1-\tau_{inc})F - v^L \geq \underline{U}\,, \tag{10}$$

$$F \geq 0. \tag{11}$$

Der Prinzipal muss hierbei lediglich die Teilnahme des Agenten sicherstellen und das Grundgehalt F ausreichend hoch wählen. Die Haftungsbeschränkung (11) ist redundant, da diese stets durch $F \geq \frac{\underline{U}+v^L}{1-\tau_{inc}}$ aus (10) sichergestellt ist. Des Weiteren folgt aus der bindenden Teilnahmebedingung unmittelbar für die Höhe des Grundgehalts: $F = \frac{\underline{U}+v^L}{1-\tau_{inc}}$. Der Prinzipal wählt nun diejenige Alternative, die durch Einsetzen der hergeleiteten Entlohnungen in die Zielfunktion sein Endvermögen maximiert, also $EV^* = \max\{EV\left(a^L\right), EV\left(a^H\right)\}$.

3.3. Steuerinduzierte Effekte bei unbeobachtbarer Arbeitsanstrengung

Ein Vorteil aus der Implementierung des hohen Anstrengungsniveaus ist bei unbeobachtbarer Handlung trotz der ex ante effizienten Anstrengung a^H durch Ungleichung (3)

[170] Angenommen $w^L > 0$. Dann könnte der Prinzipal durch Senkung von w^L bis auf 0 sein Endvermögen erhöhen, ohne eine der Nebenbedingungen zu verletzen. Damit kann nur $w^L = 0$ optimal sein.

[171] Eine formale Ausführung der First-best-Lösung erfolgt in Kapitel 3.4.

[172] Vgl. Kräkel (2007), S. 27f.

aus Prinzipalsicht nicht zwingend gegeben. Dies ist daran erkennbar, dass die ex ante Effizienz nicht automatisch zu einem ex post höheren Endvermögen dieser Alternative beim Prinzipal führt. Insbesondere kann die Anreizsetzung für den Prinzipal mit zu hohen (Agentur-)Kosten verbunden sein.

Um dies zu sehen, werden zunächst die Endvermögenshöhen der beiden Alternativen bestimmt. Für die Implementierung von a^H ergibt sich diese durch Einsetzen von $w^H = \frac{\Delta v}{\Delta \phi} \frac{1}{1-\tau_{inc}}$ und $w^L = 0$ in Gleichung (1) und beträgt damit:

$$EV\left(a^H\right) = (1-\tau_d)\Bigg[(1-\tau_c)\left(\phi^H\left(x^H - \frac{\Delta v}{\Delta\phi}\frac{1}{1-\tau_{inc}}\right) + \left(1-\phi^H\right)x^L\right) - k\tau_c\left(\phi^H\frac{\Delta v}{\Delta\phi}\frac{1}{1-\tau_{inc}}\right)\Bigg]. \tag{12}$$

Analog ergibt sich für das Endvermögen bei Zahlung des Grundgehalts $F = \frac{\underline{U}+v^L}{1-\tau_{inc}}$:

$$EV\left(a^L\right) = (1-\tau_d)\Bigg[(1-\tau_c)\left(\phi^L x^H + \left(1-\phi^L\right)x^L - \frac{\underline{U}+v^L}{1-\tau_{inc}}\right) - k\tau_c\frac{\underline{U}+v^L}{1-\tau_{inc}}\Bigg]. \tag{13}$$

Um die Steuerwirkungen explizit herauszuheben, erfolgt zunächst der Vergleich der Alternativen in einer Welt ohne Steuern. Hierzu wird der Spezialfall $EV_{\tau=0}\left(a^L\right)$ bzw. $EV_{\tau=0}\left(a^H\right)$ betrachtet. Diese Schreibweise mit $\tau = 0$ im Index steht für eine Welt ohne jegliche Steuern, sodass alle Steuersätze auf Null gesetzt werden. Der Vergleich von $EV_{\tau=0}\left(a^H\right) \gtreqless EV_{\tau=0}\left(a^L\right)$ führt auf

$$\begin{aligned} &\phi^H\left(x^H - \frac{v^H - v^L}{\phi^H - \phi^L}\right) + \left(1-\phi^H\right)x^L \gtreqless \\ &\qquad \phi^L x^H + \left(1-\phi^L\right)x^L - \underline{U} - v^L \\ \Leftrightarrow \quad &\qquad \left(\phi^H - \phi^L\right)\Delta x \gtreqless \\ &\qquad \phi^H\frac{v^H - v^L}{\phi^H - \phi^L} - \left(\underline{U} + v^L\right). \end{aligned} \tag{14}$$

Die linke Seite von (14) ist offensichtlich strikt positiv. Die rechte Seite ist ebenfalls strikt positiv, weil dies durch Ungleichung (9) und $v^H > v^L$ sichergestellt ist. Folglich ist ohne konkrete Parameterwerte nicht feststellbar, welche Alternative für den Prinzipal vor Steuern optimal ist. Damit geht insbesondere die Möglichkeit einher, dass der Fixlohnvertrag dem Prinzipal ein höheres Endvermögen liefert.

Trifft dieser Fall zu und ist die linke Seite von (14) kleiner als die rechte, so kann die ex ante Effizienzbedingung unter Umständen den zulässigen Parameterraum für diesen Fall einschränken. Für $\underline{U} = v^L = 0$ steht $EV\left(a^L\right) > EV\left(a^H\right)$ allerdings nie im Widerspruch zu der ex ante Effizienzbedingung. Demnach liegt keine weitere Einschränkung als (14) vor.[173] Diese Parameterwerte sind von Bedeutung, da hierdurch die "ungewünschte" First-best-Lösung (im nächsten Kapitel) ausgeschlossen wird. Insgesamt ist aus der vorsteuerlichen Betrachtung festzuhalten, dass die Ineffizienz aus der unbeobachtbaren Handlung nicht alleine auf Steuerwirkungen zurückzuführen ist. Der zentrale Trade-off besteht aus Sicht des Prinzipals darin, den höheren, erwarteten Cash-Flow (linke Seite) gegen die für die Implementierung von a^H höhere, erwartete Entlohnungszahlung abzuwägen.

Die Gegenüberstellung der Alternativen nach Steuern erfordert die Analyse von $EV\left(a^H\right) \gtreqless EV\left(a^L\right)$. Im Anhang[174] wird gezeigt, dass der Vorteilhaftigkeitsvergleich auf die Bedingung

$$\left(\phi^H - \phi^L\right)\Delta x \gtreqless \frac{1 + k\frac{\tau_c}{1-\tau_c}}{1-\tau_{inc}}\left(\phi^H\frac{v^H - v^L}{\phi^H - \phi^L} - \left(\underline{U} + v^L\right)\right) \tag{15}$$

führt. Die Nachsteuerbedingung ähnelt der Vorsteuerbedingung strukturell. Beide Seiten sind wiederum positiv, sodass es abhängig von den Parameterwerten vorteilhaft für den Prinzipal sein kann, die hohe oder niedrige Anstrengung zu induzieren. Der entscheidende Unterschied liegt im Faktor $\left(1 + \frac{k\tau_c}{1-\tau_c}\right)(1-\tau_{inc})^{-1}$, der größer oder gleich Eins ist. Dieser erhöht die erwarteten Entlohnungskosten, sodass es nach Steuern wahrscheinlicher ist, dass der Prinzipal den Fixlohnvertrag, nicht aber den leistungsabhängigen Vertrag wählt.

Bei genauerer Betrachtung der Bedingung (15) wird deutlich, dass hinsichtlich der Alternativenwahl die proportionale Steuer auf Dividenden τ_d keine Vorteilhaftigkeitsänderung bewirkt, zumal diese bei beiden Alternativen gleichermaßen auf das ausgeschüttete Residuum anfällt. Dies entspricht einer entscheidungsneutralen Zielgrößenbesteuerung beim Prinzipal.[175] Bei unbeschränkt abziehbarem Managergehalt $k = 0$ wäre die Bedingung ebenfalls komplett unabhängig von dem Unternehmensteuersatz τ_c, sodass der Unternehmensteuersatz zwar den Konsum des Prinzipals mindert, nicht jedoch die Wahl seiner gewünschten Alternative im Vergleich zum Vorsteuerfall verzerrt. Dies liegt daran, dass für $k = 0$ die Unternehmensteuer als Cash-Flow-Steuer ausgestaltet wäre und deshalb entscheidungsneutral ist.[176]

Bei beschränkter Abzugsfähigkeit des Managergehalts $k > 0$ steigt der Faktor im Unternehmensteuersatz τ_c. Dies hat zur Folge, dass bei höherem τ_c die Fixlohnvariante im Vergleich zur leistungsabhängigen Entlohnung begünstigt wird. Der Grund hierfür liegt darin, dass bei der leistungsabhängigen Entlohnung die zusätzliche Steuerlast aus der

[173] Für den allgemeinen Fall wird die untere Schranke $\Delta x = \frac{v^H - v^L}{\phi^H - \phi^L}$ aus der ex ante Effizienzbedingung in (14) eingesetzt. Umformen führt zu der zusätzlichen Bedingung $\underline{U}\left(\phi^H - \phi^L\right) < \phi^L v^H - \phi^H v^L$, die für die Vorteilhaftigkeit des Fixlohnvertrages erfüllt sein muss.

[174] Der Anhang wird auf der Webseite von JUMS zur Verfügung gestellt.

[175] Vgl. Wagner (1984), S. 203.

[176] Vgl. Meißner et al. (2014), S. 222.

beschränkten Abzugsfähigkeit größer ist, da wegen Ungleichung (9) die erwartete, leistungsabhängige Entlohnung mit $\phi^H w^H = \phi^H \frac{v^H - v^L}{\phi^H - \phi^L} \frac{1}{1-\tau_{inc}}$ diejenige bei der Zahlung eines Grundgehaltes mit $F = \frac{\underline{U}+v^L}{1-\tau_{inc}}$ übersteigt. Damit kann die beschränkte Abzugsfähigkeit im Optimum eine Reduktion des Aktivitätsniveaus bewirken und verringert den Kooperationsgewinn des Prinzipals.[177] Je restriktiver die Abzugsfähigkeit ist ($k \uparrow$), desto stärker verringert sich hierbei der Kooperationsgewinn des Prinzipals bei gegebenem Unternehmensteuersatz.

Die Einkommensteuer des Managers bewirkt eine steuerinduzierte Verzerrung, da das Anstrengungsniveau a^H durch sie benachteiligt wird. Diese Erkenntnis ist bereits mit Blick auf die konkreten Entlohnungshöhen zu erkennen, weil die nachsteuerliche Entlohnung $w^H = \frac{v^H - v^L}{\phi^H - \phi^L} \frac{1}{1-\tau_{inc}} > \frac{v^H - v^L}{\phi^H - \phi^L} = w^H_{\tau=0}$ gerade so korrigiert wird, dass der Agent netto die selbe Rente wie im Vorsteuerfall erzielt. Gleiches gilt für die Fixlohnalternative. Da der Agent bei der leistungsabhängigen Entlohnung eine höhere, erwartete Steuerlast zu tragen hat und diese komplett an den Prinzipal überwälzt, ist auch die Einkommensteuer des Managers ein negativer Einflussfaktor aus Sicht des Prinzipals. Hintergrund des Einflusses von τ_{inc} ist wiederum, dass beim Agenten keine Zielgrößenbesteuerung vorliegt.[178] Hierzu müssten die Disnutzen v^j sowie der Reservationsnutzen $\underline{U}$ ebenfalls dem Steuersatz τ_{inc} unterliegen. Dann vereinfachte sich die Bedingung (15) auf einen Ausdruck, der unabhängig von τ_{inc} ist, weil die Partizipationsbedingung des Agenten (4) unabhängig von τ_{inc} wäre.

Bis zu diesem Punkt wurde festgestellt, dass Ertragsteuern Auswirkungen auf die Zielfunktion des Prinzipals und sein Vertragsangebot an den Agenten haben. In diesem Modellrahmen mit binärem Anstrengungsniveau mindert insbesondere τ_{inc} das Endvermögen vom Prinzipal, nicht aber den Nettonutzen des Agenten. Der Agent scheint als passiver Akteur unbeeinflusst von der Einkommensteuer das vom Prinzipal induzierte Anstrengungsniveau zu wählen. Im Anhang wird gezeigt, dass in einem allgemeineren Modellrahmen mit kontinuierlicher Anstrengungswahl des Agenten der Nutzen von Prinzipal und Agent durch τ_{inc} vermindert wird. Dies liefert eine genauere Interpretation des hiesigen Modellrahmens: Auch wenn dem Anschein nach der Agent durch sämtliche Ertragsteuern nicht benachteiligt wird, so sinkt jedoch die Wahrscheinlichkeit, dass er zu einer hohen Arbeitsanstrengung durch den Prinzipal motiviert wird. Dadurch sinkt für ihn die Wahrscheinlichkeit einer Rente, sodass Steuern letztlich den Agentennutzen negativ beeinflussen.[179]

3.4. First-best-Verträge unter Berücksichtigung von Steuern

Im vorherigen Abschnitt wird gezeigt, dass Steuern das ex ante effiziente Anstrengungsniveau verhindern können, wenn die Handlung des Agenten unbeobachtbar ist. Zudem stellt sich heraus, dass der Agent eine Rente erzielt, die aus Prinzipalsicht mit Nutzeneinbußen verbunden ist. Nachfolgend wird der First-best-Vertrag unter Berücksichtigung des bisher analysierten Steuersystems als Benchmark charakterisiert. Hinsichtlich der Fixlohnalternative ist festzustellen, dass es keinen besseren als den bereits dargestellten Vertrag gibt, der das niedrige Anstrengungsniveau implementiert. Damit entspricht das Endvermögen bei unbeobachtbarer Anstrengung demjenigen im First-best-Fall (Index FB): $EV\left(a^L\right) = EV^{FB}\left(a^L\right)$. Einbußen aus der Unbeobachtbarkeit gibt es einzig beim leistungsabhängigen Vertrag.

Der Prinzipal muss im leistungsabhängigen First-best-Vertrag lediglich sicherstellen, dass die Partizipationsbedingung des Agenten erfüllt ist. Die Anreizkompatibilitätsbedingung entfällt, da der Prinzipal durch die Beobachtbarkeit der Handlung den Agenten bei Abweichung von a^H hinreichend bestrafen kann. Er bietet dem Agenten eine Entlohnung w^{FB} an, wenn dieser a^H wählt und zahlt dem Agenten bei Beobachtung von a^L keine Entlohnung. Dadurch stellt der Prinzipal sicher, dass der Agent stets die hohe Anstrengung wählt und immer w^{FB} erhält. Diese Vertragsform wird als Forcing-Vertrag bezeichnet und lässt sich formal wie folgt charakterisieren:

$$w(a) = \begin{cases} w^{FB} & , a = a^H \\ 0 & , a = a^L \end{cases} \tag{16}$$

Die bindende Teilnahmebedingung des Agenten

$$(1-\tau_{inc})\left(\phi^H w^{FB} + \left(1-\phi^H\right) w^{FB}\right) - v^H = \underline{U} \tag{17}$$

führt auf die Entlohnungshöhe $w^{FB} = \frac{v^H + \underline{U}}{1-\tau_{inc}}$. Einsetzen in die Zielfunktion des Prinzipals aus (1) ergibt

$$EV^{FB}\left(a^H\right) = (1-\tau_d)\big[(1-\tau_c) \left(\phi^H x^H + \left(1-\phi^H\right) x^L - w^{FB}\right) - k\tau_c w^{FB}\big]. \tag{18}$$

Aus der Gegenüberstellung mit dem Endvermögen bei unbeobachtbarer Anstrengung wird ersichtlich, dass $EV^{FB}\left(a^H\right) > EV\left(a^H\right)$ gilt.[180] Hierdurch wird bestätigt, dass der Prinzipal neben den gesunkenen Rückflüssen durch steuerliche Vorschriften Nutzeneinbußen durch die unbeobachtbare Handlung des Agenten hat.

Ein Vergleich der Zielfunktionswerte $EV^{FB}\left(a^H\right)$ und $EV^{FB}\left(a^L\right) = EV\left(a^L\right)$ liefert im Allgemeinen keine Dominanz eines der beiden Aktivitätsniveaus. $EV^{FB}\left(a^H\right)$ ist nur

[177] Dieses Ergebnis stimmt mit den bisherigen Ergebnissen bei einem risikoaversen Agenten überein. Vgl. Voßmerbäumer (2012), S. 549f; Bauer et al. (2018), S. 40f. Wenn ausschließlich das nicht leistungsabhängige Grundgehalt beschränkt abziehbar ist, wie es im amerikanischen Steuerrecht ab einem Freibetrag von $1.000.000 der Fall ist, so würde in dem Modell die hohe induzierte Anstrengung unmittelbar begünstigt. Das Ergebnis bliebe bei Risikoaversion des Managers bestehen. Vgl. Halperin, Kwon und Rhoades-Catanach (2001), S. 59f.

[178] Vgl. Jasper (1995), S. 90 und 93.

[179] Diese Klarstellung hinsichtlich des Einflusses auf den Agentennutzen dient dazu, die Funktionsweise des Modellrahmens bei binärer Anstrengung hervorzuheben. In Kapitel 4 wird auf den Agentnutzen nicht mehr explizit eingegangen, da eine Beurteilung der Vorteilhaftigkeit von Anreizschemata aus Eignersicht erfolgt.

[180] Siehe Anhang.

dann größer als $EV^{FB}\left(a^L\right)$,[181] wenn

$$\left(\phi^H-\phi^L\right)\Delta x > \underbrace{\left(1+\frac{k\tau_c}{1-\tau_c}\right)(1-\tau_{inc})^{-1}}_{\geq 1}\left(\nu^H-\nu^L\right). \quad (19)$$

Offensichtlich beeinflusst das Steuersystem die First-best-Lösung. Die ex ante Effizienzbedingung nach Steuern (19) ist durch den steuerinduzierten Faktor auf der rechten Seite restriktiver als die korrespondierende Vorsteuerbedingung (3), die sich durch Setzen von $\tau_c = 0$ und $\tau_{inc} = 0$ als Spezialfall ergibt. Analog zum Fall bei unbeobachtbarer Arbeitsanstrengung führen sowohl $k > 0$ mit $\tau_c > 0$ als auch $\tau_{inc} > 0$ zu einer gesunkenen Vorteilhaftigkeit der Implementierung einer hohen Leistung. Die Steuerwirkungen sind damit unabhängig von der (Nicht-)Beobachtbarkeit der Agentenhandlung qualitativ identisch.

Die Unbeobachtbarkeit der Agentenhandlung ist nicht allein dafür verantwortlich, dass für den Prinzipal das höhere Endvermögen der First-best-Lösung unerreichbar ist. Sofern die Haftungsbeschränkung beim Agenten nicht existent wäre, könnte der Prinzipal trotz unbeobachtbarer Anstrengung durch Belohnung $w^{H,FB} > 0$ bei hohem Output bzw. Verlustbeteiligung $w^{L,FB} < 0$ bei niedrigem Output die First-best-Lösung erzielen, da die Endvermögenshöhe derjenigen bei beobachtbarer Handlung entspricht.[182] Dieser in der Praxis aufgrund von arbeitsrechtlichen Hürden irrelevante Fall wurde in Übereinstimmung mit Kapitel 2.2 durch das Fehlen eines Haftungsvermögens beim Agenten ($W = 0$) ausgeschlossen.

In der einperiodigen Vertragsbeziehung wurde das fixe Grundgehalt und der periodisch ausgezahlte Bonus unabhängig von der (Nicht-)Beobachtbarkeit der Anstrengungswahl als mögliche Vergütungsform identifiziert. Im Mittelpunkt der Arbeit steht allerdings die Vorteilhaftigkeitsanalyse von Bonusbankvereinbarungen. Für eine vollständige Vorteilhaftigkeitsanalyse von Anreizsystemen wie der Bonusbank muss nach Jasper (1995) das Vergleichsobjekt bzw. die "relevante Alternative"[183] zur Bonusbank festgelegt werden. In diesem Kapitel ist deutlich geworden, dass es zwei relevante Alternativen zur Bonusbank gibt. Dies ist einerseits der periodisch ausgezahlte Bonus und andererseits das fixe Grundgehalt. Allgemeingültige Erkenntnisse sind nur unter Berücksichtigung beider Alternativen identifizierbar, zumal steuerliche Größen neben dem Einfluss auf die Bonusbank auch unterschiedliche Auswirkungen auf die beiden Alternativen haben. Daher werden das Grundgehalt und die periodische Bonuszahlung explizit im Analyseteil miteinbezogen, während in bisherigen Untersuchungen eine Beschränkung auf lediglich eine der Alternativen vorlag.[184]

In Anbetracht der im nächsten Kapitel unterstellten Bedingung $\underline{U} = \nu^L = 0$ zur Sicherstellung eines Anreizproblems sowie der ex ante Effizienzbedingung nach Steuern (19) lassen sich Eigenschaften herleiten, die für die folgende Analyse von Bedeutung sind: Zum einen ist die in Kapitel 3.1 getroffene Annahme eines stets positiven Residuums unabhängig vom Output und von der Entlohnung des Agenten für alle $x^L \geq 0$ nicht einschränkend. Bei niedrig realisiertem Output ist die Eigenschaft wegen $F = \frac{\underline{U}+\nu^L}{1-\tau_{inc}} = 0 = w^L$ stets erfüllt.[185] Zum anderen muss bei Realisation eines hohen Outputs das Residuum des Prinzipals immer positiv sein, da ansonsten der irrelevante Fall eintritt, dass die hohe Anstrengung nie durch den Prinzipal implementiert wird. Dann wäre ein Fixlohnvertrag für den Prinzipal immer optimal.[186] Insgesamt ist durch den geschaffenen Modellrahmen nun eine präzise Vorteilhaftigkeitsanalyse der Bonusbank möglich. Diese wird im nächsten Kapitel vorgenommen.

4. Vorteilhaftigkeitsanalyse einer Bonusbank

Zur Analyse der Bonusbank ist eine mehrperiodige Betrachtung der Prinzipal-Agenten-Beziehung erforderlich. Deshalb wird in diesem Kapitel zunächst auf einige Punkte eingegangen, die sich konzeptionell aus einer mehrperiodigen Erweiterung ergeben. Daraus resultiert eine erste Prämissensetzung. Anschließend erfolgt die formale Vorteilhaftigkeitsuntersuchung der Bonusbank mit besonderem Fokus auf den Einfluss von steuerlichen Gegebenheiten. Steuerliche Verluste auf Unternehmensebene werden in einem Teil der Analyse wie in Kapitel 3 ausgeschlossen, in einem anderen Teil jedoch explizit in die Analyse integriert. Abgeschlossen wird die Vorteilhaftigkeitsanalyse mit einer kritischen Würdigung der Annahmen, die für die nachfolgende Untersuchung notwendig sind.

4.1. Aspekte einer mehrperiodigen Erweiterung

4.1.1. Umgang mit der Mehrperiodigkeit

Die ökonomische Wirklichkeit ist typischerweise durch langfristige Vertragsbeziehungen zwischen Prinzipal und Agent sowie eine wiederholte Aufgabendurchführung gekennzeichnet.[187] Angesichts der mittlerweile stark erforschten, steuerlichen Anreizwirkungen in einperiodigen Prinzipal-Agenten-Modellen ist es verwunderlich, dass bisher noch keine mehrperiodige Erweiterung unter Berücksichtigung steuerlicher Aspekte vorgenommen wurde. Während in einperiodigen Prinzipal-Agenten-Beziehungen nur Bemessungs-

[181] Siehe Anhang.
[182] Siehe Anhang.
[183] Jasper (1995), S. 122.
[184] Bei der Untersuchung von verzögerten Auszahlungsplänen unter Vernachlässigung ihrer Anreizwirkung ist als Alternative das Fixum genutzt worden. Vgl. Miller und Scholes (1982), S. 184; Jasper (1995), S. 122. In der einzigen Untersuchung einer Bonusbank, allerdings unter Vernachlässigung von Steuern, nutzte Janocha den periodischen Bonus als alleinige Alternative. Vgl. Janocha (2014), S. 89ff.
[185] Einsetzen der Entlohnungen ergibt: $x^L \geq (1+\frac{k\tau_c}{1-\tau_c})w^L = 0$ und $x^L \geq (1+\frac{k\tau_c}{1-\tau_c})F = 0$.
[186] Siehe Anhang.
[187] Vgl. Schäfer (2013), S. 129.

grundlagen- und Steuertarifeffekte aufgezeigt werden können, sind Zeit- bzw. Zinseffekte nicht abbildbar.[188] Zur Analyse langfristiger Anreizsysteme sind diese aber, wie sich zeigen wird, von zentraler Bedeutung.

In einer mehrperiodigen Betrachtung sind zusätzliche Prämissen notwendig, um die Analyse der steuerlichen Wirkungen hinsichtlich der Vorteilhaftigkeit von Bonusbankvereinbarungen zu ermöglichen. Aus Vereinfachungsgründen beschränkt sich die Analyse im Folgenden auf eine zweiperiodige Vertragsbeziehung, auch wenn Bonusbankvereinbarungen im Regelfall über einen längeren als den Zwei-Jahres-Zeitraum geschlossen werden. Zur Darstellung der wesentlichen Eigenschaften einer Bonusbank ist dies allerdings ausreichend. Der Prinzipal bietet dem Agenten einen langfristigen Vertrag über die gesamten zwei Perioden an, der rechtlich bindend ist und nicht nachverhandelt wird.[189] In dem Vertrag ist geregelt, ob der Prinzipal eine Bonusbank, einen periodisch ausgezahlten Bonus oder ein Grundgehalt nutzt. Der Prinzipal hat nicht die Möglichkeit, den Agenten zu entlassen, weil dies bspw. mit zu hohen Entschädigungszahlungen einhergeht. Der Agent kündigt seinerseits nicht, da der Vertrag ihm einen entsprechend hohen, intertemporalen Nutzen liefert.

In Zusammenhang mit einer mehrperiodigen Erweiterung ist eine Bewertung von intertemporalen Zahlungsströmen sowohl beim Prinzipal als auch beim Agenten abhängig von einem Kapitalmarktzugang.[190] Die Gewinnerzielung beim Prinzipal und die Entlohnung beim Agenten liefert keinen Nutzen an sich. Der Nutzen der Akteure hängt von den Folgemaßnahmen der Periode in Gestalt der Konsummöglichkeiten ab.[191] Durch einen Kapitalmarkt können die Akteure Zahlungsströme so verschieben, dass dies mit Blick auf ihre intertemporale Konsumfunktion optimal ist. Allerdings erschwert dies die Bewertung von Anreizverträgen, da Konsum- und Sparentscheidungen letztlich ein weiteres Optimierungsproblem der Akteure darstellt.

Vereinfachend wird daher unterstellt, dass beim Prinzipal und beim Agenten keine Kapitalmarktaktivitäten durchgeführt werden. Auf Prinzipalebene kann das dadurch gerechtfertigt werden, dass dieser im Sinne einer langfristigen Unternehmenswertsteigerung die Konsummöglichkeiten zum Ende des Kooperationszeitraums maximiert. Das maximale Endvermögen auf Ebene des repräsentativen Gesellschafters hängt dann bei Konstanthaltung des zu bestimmenden, optimalen Vergütungssystems ausschließlich von der Ausschüttungspolitik ab. Diese ist entweder die periodische Ausschüttung sämtlicher Zahlungsüberschüsse an den Eigner, der die freien Mittel im Privatvermögen anlegt, oder die Investition frei gewordener Überschüsse am Kapitalmarkt durch die Kapitalgesellschaft.

Bei Unterstellung eines einheitlichen Bruttozinssatzes i ist die optimale Ausschüttungspolitik lediglich abhängig von den vorliegenden Steuersätzen im Privat- bzw. Betriebsvermögen.[192] Unter Zugrundelegung der Tatsache, dass Zinseinkünfte in Zukunft nicht mehr der Abgeltungsteuer, sondern der tariflichen Einkommensteuer unterliegen sollen, kann die Anlage auf Unternehmensebene als Ergebnis einer rationalen Ausschüttungspolitik gewertet werden.[193] Wenn die Kapitalgesellschaft die freien Mittel in eigene Projekte investiert, deren Rendite größer ist als bei einer einfachen Kapitalmarktanlage, so ist dies auch für die aktuelle Rechtslage zu rechtfertigen, falls die Nachsteuerrendite größer ist als der Nachsteuerzins bei Anlage im Privatvermögen.

Da in der Arbeit eine Beurteilung der Anreizschemata aus Prinzipalsicht erfolgt, erscheint die Vernachlässigung von Kapitalmarktaktivitäten des Agenten wenig restriktiv. Allerdings sind die Kapitalmarktaktivitäten des Agenten ebenfalls für den Prinzipal von Bedeutung, da Sparentscheidungen des Agenten in früheren Perioden negative Anreizwirkungen in zukünftigen Perioden verursachen können.[194] In der Arbeit wird dem dadurch Rechnung getragen, dass der Agent die an ihn ausbezahlte Entlohnung sofort in der Verdienstperiode konsumieren muss.[195] Gleichzeitig besitzt der Agent eine kurzfristige Konsumpräferenz, sodass er den sofortigen Konsum einem Konsum in späteren Perioden vorzieht.

4.1.2. Formalisierung des Mehrperiodenfalls

Nach Abschluss des langfristigen Vertrages über den Kooperationszeitraum von zwei Perioden wählt der Agent in jeder Periode die unbeobachtbare Arbeitsanstrengung a_t^j mit $j \in \{L, H\}$, die den Cash-Flow x_t^j mit $x_t^H > x_t^L \geq 0$ der jeweiligen Periode $t \in \{1, 2\}$ beeinflusst. Anders als im einperiodigen Modell wird ausschließlich der Second-best-Fall untersucht.[196] Die Realisation des Outputs bestimmt die verdienten Entlohnungen des Agenten und ist abhängig von der gewählten Vergütungsform des Prinzipals. Zum Ende der Periode t zahlt der Prinzipal dem Agenten die Entlohnungen aus, sodass die Steuerzahlungen bei der Kapitalgesellschaft mit Steuersatz τ_c und beim Agenten mit Steuersatz $\tau_{inc,t}$ fällig werden. Sämtliche Überschüsse, die zum Ende der ersten Periode vorliegen, werden von der Kapitalgesellschaft auf dem KM zum (Brutto-)Zinssatz $i \geq 0$ angelegt und stellen steuerbare Einkünfte in der zweiten Periode dar. Zum Ende der zweiten Periode werden alle Überschüsse des Unternehmens

[188] Die Systematisierung der steuerlichen Wirkungen in diese drei Effekte geht auf Wagner (1984) zurück.

[189] Bierbaum (2002) stellt fest, dass langfristige Verträge nicht immer durch den Prinzipal gewählt werden. Insofern wird davon ausgegangen, dass kurzfristige Verträge mit anderweitigen Nachteilen verbunden sind, die außerhalb des Modellrahmens liegen. Schöndube (2006) thematisiert Nachverhandlungen in langfristigen Anreizbeziehungen.

[190] Vgl. Schäfer (2013), S. 133.

[191] Vgl. Laux (2003), S. 37; König und Wosnitza (2004), S. 78.

[192] Vgl. Kiesewetter und Dietrich (2007), S. 237.

[193] Vgl. CDU/CSU und SPD (2018), S. 69. Wird ein Grenzeinkommensteuersatz von 42% beim Prinzipal unterstellt, so ist eine Anlage im Privatvermögen unter Berücksichtigung des Solidaritätszuschlags nur für unrealistische hohe Gewerbesteuerhebesätze von über 813% optimal.

[194] Vgl. Dierkes und Schäfer (2008), S. 239.

[195] Diese Annahme ist in mehrperiodigen Prinzipal-Agenten-Modellen üblich. Vgl. z. B. Lambert (1983), S. 449.

[196] Der First-best-Fall wird mit Verweis auf die analoge Ermittlung in Kapitel 3.4 nicht explizit dargestellt.

an den Anteilseigner ausgeschüttet, sodass dieser das nachsteuerliche Endvermögen für Konsumzwecke zur Verfügung hat. Abbildung 5 veranschaulicht die beschriebene Zeitabfolge der mehrperiodigen Erweiterung.

Wie im einperiodigen Modell ist der Zusammenhang zwischen Output und Anstrengungsniveau in der Periode t durch $Pr(\tilde{x}_t = x_t^H | a_t = a_t^j) = \phi_t^j$ gegeben, wobei zum Ausschluss trivialer Lösungen $\phi_t^j \in (0,1)$ und $\phi_t^H > \phi_t^L$ erfüllt sind. Die gegebenen Wahrscheinlichkeiten sind auf gegebene Investitionsentscheidungen zurückzuführen. Dadurch sind Abschreibungsbeträge in jeder Periode deterministisch und werden vernachlässigt.[197] Die einzelnen Perioden sind unabhängig voneinander, sodass jegliche Anstrengung einer Periode den Cash-Flow einer anderen Periode unberührt lässt. Sei $m \in \{L, H\}$ eine weitere Indexvariable für die Ausprägung hoch oder niedrig. Dann lässt sich die Periodenunabhängigkeit formal folgendermaßen ausdrücken:

$$Pr(\tilde{x}_1, \tilde{x}_2 | a_1^j, a_2^m) = Pr(\tilde{x}_1 | a_1^j)\, Pr(\tilde{x}_2 | a_2^m), \qquad \forall j, m. \quad (20)$$

Damit können in dem Mehrperiodenfall vier Zustände (x_1^j, x_2^m) mit positiver Wahrscheinlichkeit eintreten, wobei sich die Eintrittswahrscheinlichkeit durch Multiplikation der periodischen Wahrscheinlichkeiten ergibt.

Der Prinzipal hat drei mögliche Vergütungssysteme zur Auswahl. Zum einen kann er dem Agenten ein Grundgehalt F zum Ende der Perioden zahlen, welches das niedrige Anstrengungsniveau in beiden Perioden implementiert. Zum anderen kann er dem Agenten erfolgsabhängige Entlohnungen $\tilde{w}_t \in \{w_t^H, w_t^L\}$ komplett in der Verdienstperiode auszahlen (periodische Vergütung), um in jeder Periode eine hohe Anstrengung des Agenten zu induzieren. Dies sind die in Kapitel 3.4 beschriebenen relevanten Alternativen zur Bonusbank. Sie liefern dem Prinzipal ein Endvermögen von $EV(a_1^L, a_2^L) = EV(a^L)$ bzw. $EV(a_1^H, a_2^H) = EV(a^H)$. Wohlfahrtsoptimal ist in jeder Periode das hohe Anstrengungsniveau, da die ex ante Efizienzbedingung für jede Periode durch $\Delta x_t \Delta \phi_t > \Delta v_t$ erfüllt ist.

Alternativ dazu kann der Prinzipal zur Implementierung des hohen Anstrengungsniveaus in beiden Perioden eine Bonusbankvereinbarung nutzen.[198] Die Bonusbank wird wie in Janocha (2014) modelliert.[199] Bei einer Bonusbank (hochgestellter Index BB) verdient der Manager in jeder Periode eine Entlohnung $w_t^{j,BB}$ in Abhängigkeit vom realisierten Output. Allerdings wird der verdiente Bonus aus der ersten Periode nicht wie bei einer periodischen Vergütung komplett ausbezahlt. Ein Anteil $\alpha \in [0,1)$ des in der ersten Periode verdienten Bonus $w_1^{j,BB}$ wird in die Bonusbank eingestellt und nur ausgezahlt, wenn in der zweiten Periode ein hoher Cash-Flow realisiert wird. Andernfalls entfällt das Bonusbankguthaben aus der ersten Periode.

In der Tabelle 2 werden die resultierenden Unterschiede in den gezahlten Entlohnungen bei einer periodischen Vergütung und einer Bonusbank aufgezeigt. Von einem Startguthaben wird ebenso wie von zusätzlichen Kosten der Bonusbankeinführung abgesehen. Wählt der Prinzipal die Bonusbank als Anreizsystem, so erzielt er ein Endvermögen $EV^{BB}(a^H)$, das sich durch Multiplikation der Wahrscheinlichkeiten mit den zustandsabhängigen Endvermögen ergibt. Die Wahrscheinlichkeiten sind bei einer periodischen Vergütung und einer Bonusbank identisch, da jeweils das hohe Anstrengungsniveau motiviert wird. Im Optimum wählt der Prinzipal die Alternative mit dem höchsten Endvermögen: $EV^* = \max\{EV\left(a^L\right), EV\left(a^H\right), EV^{BB}\left(a^H\right)\}$.

In den Entlohnungszahlungen muss der Prinzipal berücksichtigen, dass der Agent einen periodenunabhängigen Disnutzen aus der Wahl des hohen Anstrengungsniveaus $v_t^H = v^H > 0$ erleidet. Des Weiteren wird übereinstimmend mit den Erkenntnissen aus dem einperiodigen Fall zur Sicherstellung eines Anreizproblems $\underline{U}_t = \underline{U} = 0 = v_t^L = v^L$ unterstellt. Ferner präferiert der Agent einen frühzeitigen Konsum gegenüber einem späteren Konsum. Daher bewertet er in der zweiten Periode erhaltene Nettoentlohnungen mit einem Diskontfaktor δ mit $0 < \delta \leq 1$. Diese Konsumpräferenzrate ist dem Prinzipal vor Vertragsabschluss bekannt. Eine Verzinsung des Bonusbankguthabens erfolgt nicht, weil Zinsverluste aus einer verzögerten Auszahlung implizit in die Höhe der Konsumpräferenzrate einfließen.

Die Kapitalgesellschaft muss einen periodenunabhängigen Unternehmensteuersatz τ_c auf ihre steuerlichen Gewinne zahlen, da sie während des Betrachtungszeitraums ihren Unternehmenssitz nicht verlagert und s^{KSt} konstant bleibt. Entlohnungszahlungen sind weiterhin in Höhe der Auszahlung anteilig $(1-k)$ als Betriebsausgabe abzugsfähig. Eine Rückstellungsbildung für den einbehaltenen Anteil $\alpha w_1^{j,BB}$ ist wegen der fehlenden Abgeltung von vergangenen Leistungen nicht erlaubt, zumal die Auszahlung an die Bedingung $\tilde{x}_2 = x_2^H$ geknüpft ist. Dies schließt auch eine Anwendung der begünstigten Besteuerung des § 34 EStG beim Manager aus. Dieser hat in der Periode t die ihm zufließende Entlohnung der tariflichen Einkommensteuer mit einem Grenzsteuersatz von $\tau_{inc,t}$ zu unterwerfen.

4.2. Das mehrperiodige Grundmodell

4.2.1. Entlohnungs- und Endvermögensbestimmung bei periodischer Vergütung und Grundgehalt

Das mehrperiodige Grundmodell zeichnet sich dadurch aus, dass analog zu Kapitel 3 in keiner der beiden Perioden unabhängig von dem Anreizsystem steuerliche Verluste auftreten und positive Ausschüttungsbeträge gesichert sind.[200]

[197] Steuerliche Wirkungen könnten durch deterministische Abschreibungen trotzdem entstehen, wenn bspw. gerade die Abschreibungsbeträge zu einem steuerlichen Verlust auf Unternehmensebene führen. Insofern werden derartige Wirkungen ebenfalls ausgeschlossen.

[198] Die Motivation der hohen Anstrengung in beiden Perioden wird als "permanente Anstrengung" des Agenten bezeichnet.

[199] Vgl. Janocha (2014), S. 91.

[200] Ähnlich zu den Bedingungen in Kapitel 3.1 sind die Voraussetzungen dafür durch $x_1^H \geq \left(1 + \frac{k\tau_c}{1-\tau_c}\right) w_1^H$ und $x_2^H \geq \left(\alpha w_1^{H,BB} + w_2^{H,BB}\right)\left(1 + \frac{k\tau_c}{1-\tau_c}\right)$ gegeben.

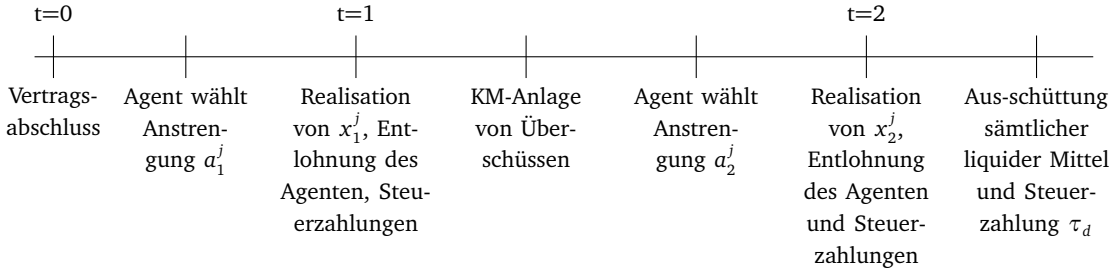


Abbildung 5: Zeitliche Abfolge im mehrperiodigen Analyseteil (Eigene Darstellung)

Zustand	Wahrschein-lichkeit	Entlohnung (periodisch)	Entlohnung (Bonusbank)	Endvermögen
(x_1^H, x_2^H)	$\phi_1^H \phi_2^H$	w_1^H, w_2^H	$(1-\alpha)w_1^{H,BB}$, $\alpha w_1^{H,BB} + w_2^{H,BB}$	EV^{HH} / $EV^{HH,BB}$
(x_1^H, x_2^L)	$\phi_1^H(1-\phi_2^H)$	w_1^H, w_2^L	$(1-\alpha)w_1^{H,BB}, w_2^{H,BB}$	EV^{HL} / $EV^{HL,BB}$
(x_1^L, x_2^H)	$(1-\phi_1^H)\phi_2^H$	w_1^L, w_2^H	$(1-\alpha)w_1^{L,BB}$, $\alpha w_1^{L,BB} + w_2^{H,BB}$	EV^{LH} / $EV^{LH,BB}$
(x_1^L, x_2^L)	$(1-\phi_1^H)(1-\phi_2^H)$	w_1^L, w_2^L	$(1-\alpha)\, w_1^{L,BB}, w_2^{L,BB}$	EV^{LL} / $EV^{LL,BB}$

Tabelle 2: Übersicht und Notation bei hoher induzierter Anstrengung im Vergleich (Eigene Darstellung)

Des Weiteren sind im Grundmodell zur Vereinfachung $\phi_t^H = \phi^H, \phi_t^L = \phi^L$ und $\tau_{inc,t} = \tau_{inc}$ periodenunabhängig.

Bei der periodischen Vergütung gilt es den kostengünstigsten Vertrag zu ermitteln, der die permanente Anstrengung (a_1^H, a_2^H) sicherstellt. Bei unbeobachtbarer Arbeitsanstrengung ist die Teilnahmebedingung des Agenten nicht bindend, da dies durch $\underline{U} = v^L = 0$ für jede einzelne Periode gewährleistet ist. Deshalb reicht zur Vertragsermittlung der Blick auf die Anreizbedingungen der Perioden. Keine Abweichung in der ersten Periode bedingt

$$\begin{aligned} &(1-\tau_{inc})\left(\phi^H w_1^H + \left(1-\phi^H\right) w_1^L\right) - v^H \geq \\ &(1-\tau_{inc})\left(\phi^L w_1^H + \left(1-\phi^L\right) w_1^L\right) \\ \Leftrightarrow \quad &\Delta w_1 = w_1^H - w_1^L \geq \frac{v^H}{\phi^H - \phi^L} \frac{1}{1-\tau_{inc}} . \end{aligned} \tag{21}$$

Keine Abweichung in der zweiten Periode führt analog zu

$$\begin{aligned} &(1-\tau_{inc})\,\delta\left(\phi^H w_2^H + \left(1-\phi^H\right) w_2^L\right) - v^H \geq \\ &(1-\tau_{inc})\,\delta\left(\phi^L w_2^H + \left(1-\phi^L\right) w_2^L\right) \\ \Leftrightarrow \quad &\Delta w_2 = w_2^H - w_2^L \geq \frac{v^H}{\phi^H - \phi^L} \frac{1}{1-\tau_{inc}} \frac{1}{\delta} . \end{aligned} \tag{22}$$

Strukturell entsprechen die Anreizbedingungen denjenigen aus dem einperiodigen Fall, weil die Entlohnungszahlungen der Perioden unabhängig voneinander sind. Dagegen wird aus der Bedingung (22) ersichtlich, dass die nötige Entlohnungsspanne zur Sicherstellung von a_2^H durch den Faktor $\delta^{-1} \geq 1$ erhöht wird. Der Agent muss für den selben Konsumnutzen in $t = 2$ härter arbeiten. Die Kosten trägt jedoch letztlich der Prinzipal, der zur Anreizsetzung die Entlohnungsspanne vergrößern muss.

Weiterhin unterliegt der Agent einer beschränkten Haftung mit Haftungsvermögen $W = 0$ in jeder Periode:

$$w_t^j \geq 0, \qquad \forall t \in \{1,2\}, j \in \{L,H\}. \tag{23}$$

Da eine Entlohnung bei niedrigem Output keine Anreizwirkung entfaltet, folgt aus der beschränkten Haftung (23) unmittelbar $w_1^L = w_2^L = 0$. Bei Indifferenz wählt der Agent stets die hohe Anstrengung. Dadurch ergeben sich die kostenminimalen Entlohnungen bei hohem Output aus den Anreizbedingungen (21) und (22) durch $w_1^H = \frac{v^H}{\phi^H - \phi^L} \frac{1}{1-\tau_{inc}}$ und $w_2^H = w_1^H \delta^{-1}$. Im Anhang wird gezeigt, dass für diese Entlohnungsgrößen die intertemporale Partizipationsbedingung nicht bindend ist, weil $(1-\tau_{inc})\,\phi^H\left(w_1^H + \delta w_2^H\right) > 2v^H$ gilt. Der Konsumnutzen aus den nachsteuerlichen Vergütungen beider Perioden übersteigt das Arbeitsleid beider Perioden. Folglich erzielt der Agent eine Rente, die den Agenturkosten aus der Anreizsetzung gleicht.[201]

Nun kann das Endvermögen des Prinzipals bestimmt werden. Hierzu wird zunächst das Endvermögen eines Zustands (x_1^j, x_2^m) mit Hilfe des Finanzplans der Tabelle 3 bestimmt. Der KM-Anlagebetrag der ersten Periode in Position [8] entspricht dem ausschüttungsfähigen Betrag aus dem einperiodigen Modell. Der erwirtschaftete Bruttozins des Anlagebetrages unterliegt in $t = 2$ dem Unternehmensteuersatz. Der ausschüttungsfähige Betrag $EV^{jm}\left(a^H\right)$ (Position [9]) am Ende von $t = 2$ ergibt sich letztlich durch Bestimmung der Nettozahlungsüberschüsse ([1]-[5]-[6]+[7]). Sei $i_{\tau_c} = i\,(1-\tau_c)$

[201] Insbesondere entspricht die Rente in jeder Periode derjenigen aus dem einperiodigen Fall repräsentiert durch Bedingung (9).

Periode		t=1	t=2
Ermittlung der Steuerlast			
[1]	Cash-Flow	x_1^j	x_2^m
[2]	Entlohnung anteilig	$-(1-k)\,w_1^j$	$-(1-k)\,w_2^m$
[3]	KM-Zinsen	0	$+i((1-\tau_c)(x_1^j-w_1^j)$ $-\tau_c k w_1^j)$
[4]	Gewerbeertrag = zvE	$=x_1^j-(1-k)\,w_1^j$	$=x_2^m-(1-k)\,w_2^m$ $+i((1-\tau_c)(x_1^j-w_1^j)-\tau_c k w_1^j)$
[5]	Unternehmensteuer	$\tau_c(x_1^j-(1-k)\,w_1^j)$	$\tau_c[x_2^m-(1-k)\,w_2^m$ $+i((1-\tau_c)(x_1^j-w_1^j)-\tau_c k w_1^j)]$
Ermittlung der Überschüsse			
[6]	Entlohnung	w_1^j	w_2^m
[7]	Habenzins und Tilgung	0	$(1+i)[(1-\tau_c)(x_1^j-w_1^j)$ $-\tau_c k w_1^j]$
[8]	KM-Anlagebetrag	$(1-\tau_c)(x_1^j-w_1^j)$ $-\tau_c k w_1^j$	0
[9]	Saldo/Endvermögen	0	$EV^{jm}\left(a^H\right)$

Tabelle 3: Finanzplan bei periodischer Vergütung und permanenter Anstrengung im Zustand (x_1^j, x_2^m) unter Ausschluss von Verlusten (Eigene Darstellung)

der erwirtschaftete Nettozins der Kapitalgesellschaft. Dann beträgt das zustandsabhängige Endvermögen vor Ausschüttung an den Gesellschafter

$$EV^{jm}\left(a^H\right)=(1-\tau_c)\left(x_2^m-w_2^m\right)-k\tau_c w_2^m+\left(1+i_{\tau_c}\right)\left((1-\tau_c)\left(x_1^j-w_1^j\right)-k\tau_c w_1^j\right). \quad (24)$$

Die erste Zeile von (24) entspricht dem erwirtschafteten Überschuss der zweiten Periode im Zustand m. Die zweite Zeile entspricht dem erwirtschafteten Überschuss der ersten Periode im Zustand j zzgl. der Nettoverzinsung durch die Kapitalmarktanlage. Eine Gewichtung der zustandsabhängigen Endvermögen mit den Eintrittswahrscheinlichkeiten

$$EV\left(a^H\right)=(1-\tau_d)\Big[\left(\phi^H\right)^2 EV^{HH}+\phi^H\left(1-\phi^H\right)EV^{HL}+\left(1-\phi^H\right)\phi^H EV^{LH}+\left(1-\phi^H\right)^2 EV^{LL}\Big] \quad (25)$$

führt auf ein Endvermögen nach Ausschüttung (siehe Anhang) in Höhe von

$$EV\left(a^H\right)=(1-\tau_d)\Big[(1-\tau_c)\left(\left(1+i_{\tau_c}\right)E(\tilde{x}_1)+E(\tilde{x}_2)\right)-\phi^H(1-\tau_c(1-k))\left(\left(1+i_{\tau_c}\right)w_1^H+w_2^H\right)\Big]. \quad (26)$$

Der Prinzipal erhält den erwarteten, nachsteuerlichen Output jeder Periode. Dabei wird der erwartete Cash-Flow der ersten Periode durch die zusätzliche Verzinsung auf den Endzeitpunkt der Kooperation bezogen. Die Entlohnungszahlungen in jeder Periode treten mit Wahrscheinlichkeit ϕ^H auf und mindern das Endvermögen. Allerdings erfolgt eine steuerliche Entlastung auf Unternehmensebene in Höhe von $\tau_c(1-k)$. Zudem geht eine höhere Entlohnungszahlung in der ersten Periode mit einem Zinsverlust einher.

Bei der Bestimmung des bestmöglichen Fixlohnvertrages zur Implementierung von a^L in beiden Perioden ist lediglich die Teilnahmebedingung zu beachten. Aus der bindenden Teilnahmebedingung

$$(1-\tau_{inc})\left(\phi^L(F+\delta F)+\left(1-\phi^L\right)(F+\delta F)\right)=(1-\tau_{inc})(1+\delta)F\overset{!}{=}0 \quad (27)$$

resultiert die Entlohnung $F=0$ im Optimum. Daraus ergibt sich ein Endvermögen in Höhe von

$$EV\left(a^L\right)=(1-\tau_d)(1-\tau_c)\left(\left(1+i_{\tau_c}\right)E(\tilde{x}_1)+E(\tilde{x}_2)\right) \quad (28)$$

mit $E(\tilde{x}_t)=\phi^L x_t^H+\left(1-\phi^L\right)x_t^L$. Damit sind die Entgelte und die Endvermögenshöhen der beiden Alternativen hergeleitet. Im Folgenden werden die konkreten Größen für die Bonusbankvereinbarung bestimmt.

4.2.2. Entlohnungs- und Endvermögensbestimmung bei Bonusbankvereinbarung

Die Bestimmung der optimalen Verträge bei permanenter Anstrengung des Agenten ist bei einer Bonusbankvereinbarung schwieriger als bei einer periodischen Vergütung. Dies liegt daran, dass die an den Agenten gezahlten Entlohnungen durch den einbehaltenen Anteil α abhängig vom realisierten Zustand sind. Dadurch können die beiden Perioden anders als bei einer periodischen Bonusausschüttung nicht mehr unabhängig voneinander betrachtet werden. Der Prinzipal muss sicherstellen, dass der Agent unabhängig vom realisierten Output der ersten Periode über den gesamten Kooperationszeitraum nicht von der Strategie (a_1^H, a_2^H) abweicht

(Ex ante Anreizkompatibilität).[202] Hierzu sind drei Nebenbedingungen ausreichend.[203] Diese werden im Anhang hergeleitet und hiernach aufgezeigt.

Der Agent wählt die hohe Anstrengung in der ersten Periode, sofern

$$\Delta w_1^{BB} = w_1^{H,BB} - w_1^{L,BB} \geq \frac{v^H}{\phi^H - \phi^L} \frac{1}{1-\tau_{inc}} \frac{1}{1-\alpha+\phi^H\delta\alpha} . \tag{29}$$

Die Bedingung (29) ist ähnlich, für $\alpha = 0$ sogar identisch zu der Bedingung (21) bei einer periodischen Vergütung. Der Unterschied für $\alpha > 0$ liegt im Faktor $\left(1-\alpha+\phi^H\delta\alpha\right)^{-1} > 1$. Der Grund hierfür liegt darin, dass die Erhöhung von α durch eine Vergrößerung der Entlohnungsspanne Δw_1^{BB} kompensiert werden muss. Ansonsten würden keine Anreize zu hoher Anstrengung in $t = 1$ gesetzt, da der ausbezahlte Betrag $(1-\alpha)w_1^{j,BB}$ im Erwartungswert zu niedrig wäre, um das entstandene Arbeitsleid zu decken. Außerdem steigt der Faktor, je kleiner die Konsumpräferenz δ des Agenten ist. Dies spiegelt den direkten Nachteil der nur anteiligen Auszahlung wider, weil der Agent den früheren Konsum präferiert.

Der Agent muss sich unabhängig von dem Output der ersten Periode in $t = 2$ anstrengen. Wenn die erste Periode ein Erfolg war ($\tilde{x}_1 = x_1^H$), lautet die Anreizbedingung $t = 2$

$$\Delta w_2^{BB} = w_2^{H,BB} - w_2^{L,BB} \geq \frac{v^H}{\phi^H - \phi^L} \frac{1}{1-\tau_{inc}} \frac{1}{\delta} - \alpha w_1^{H,BB} . \tag{30}$$

Bei Misserfolg ($\tilde{x}_1 = x_1^L$) gilt analog

$$\Delta w_2^{BB} = w_2^{H,BB} - w_2^{L,BB} \geq \frac{v^H}{\phi^H - \phi^L} \frac{1}{1-\tau_{inc}} \frac{1}{\delta} - \alpha w_1^{L,BB} . \tag{31}$$

Die Anreizbedingungen (30) und (31) für die zweite Periode zeigen, dass die Entlohnungsspanne zur Anreizsetzung abhängig vom realisierten Zustand um $\alpha w_1^{j,BB}$ geringer ist als bei einer periodischen Vergütung. Damit steht dem Nachteil der verzögerten Auszahlung durch den Faktor $\left(1-\alpha+\phi^H\delta\alpha\right)^{-1}$ dieser Vorteil gegenüber. Dem Agenten ist bewusst, dass ihm der hohe Cash-Flow die Auszahlung des einbehaltenen Bonusbankguthabens liefert. Deshalb muss der Prinzipal in $t = 2$ die Entlohnungsanreize nicht so hoch ansetzen.

Weiterhin setzt die beschränkte Haftung

$$w_t^{j,BB} \geq 0, \qquad \forall t \in \{1,2\}, j \in \{L,H\} \tag{32}$$

voraus. Aus den Bedingungen (30) bis (32) ist ersichtlich, dass $w_2^{L,BB} = 0$ die optimale Entlohnungssetzung darstellt, zumal von dieser Entlohnung keine Anreizwirkung in der ersten und zweiten Periode ausgeht. Zudem bindet wegen $w_1^{H,BB} > w_1^{L,BB}$ die Bedingung (30) nicht, während (31) im Optimum mit Gleichheit erfüllt ist. Zur Festlegung der übrigen Entlohnungen ist die Höhe von $w_1^{L,BB}$ entscheidend, da diese einen Einfluss auf alle Anreizbedingungen hat. Allerdings wird deutlich, dass der Nachteil von $w_1^{L,BB} > 0$ auf die erhöhte Entlohnungsspanne in $t = 1$ stets größer als der entstehende Vorteil in $t = 2$ ist. Während die Erhöhung von $w_1^{L,BB} > 0$ um eine Geldeinheit die Entlohnung in $t = 1$ unabhängig von dem realisierten Zustand um diese Einheit erhöht, sinkt die Entlohnung $w_2^{H,BB}$ nur um α Geldeinheiten.[204] Damit ist $w_1^{L,BB} = 0$ optimal. Die kostengünstigsten Entlohnungen bei hohem Cash-Flow lauten demnach $w_1^{H,BB} = \frac{v^H}{\phi^H-\phi^L} \frac{1}{1-\tau_{inc}} \frac{1}{1-\alpha+\phi^H\delta\alpha}$ und $w_2^{H,BB} = \frac{v^H}{\phi^H-\phi^L} \frac{1}{1-\tau_{inc}} \frac{1}{\delta}$.

Bei Annahme dieses Bonusbankvertrages erzielt der Agent über den gesamten Zeitraum dieselbe Rente wie bei einer periodischen Vergütung.[205] Dadurch wird deutlich, dass die Bonusbank den Agenten nicht an negativen Unternehmensentwicklungen teilhaben lässt. Ursächlich hierfür sind die Anreizbedingungen der zweiten Periode. Der Prinzipal setzt zur Induzierung der hohen Anstrengung bei $\tilde{x}_1 = x_1^L$ das Entgelt $w_2^{H,BB}$ so hoch, dass der Agent bei $\tilde{x}_1 = x_1^H$ eine Rente erzielt. Dadurch schwindet aber der Vorteil der doppelt leistungsabhängigen Auszahlung $\alpha w_1^{j,BB}$.

Die Entlohnungszahlungen bei einer periodischen Vergütung und einer Bonusbank werden steuerlich gleich behandelt. Endvermögensänderungen sind damit auf die unterschiedlichen Entlohnungszahlungen der Perioden zurückzuführen. Beim Vergleich dieser ist auffällig, dass die Entlohnungen bei niedrigem Cash-Flow identisch sind: $w_t^{L,BB} = w_t^L$. Außerdem wird unabhängig von $\tilde{x}_1$ in der zweiten Periode dieselbe Entlohnung gezahlt, um die Leistungserbringung des Agenten aufrechtzuerhalten: $w_2^{H,BB} = w_2^H$. Dies bedeutet, dass die Endvermögenshöhen sich im Zustand (x_1^L, x_2^m) entsprechen: $EV^{Lm}\left(a^H\right) = EV^{Lm,BB}$. In den anderen beiden Zuständen unterscheiden sich die Endvermögenshöhen. Einsetzen der Entlohnungszahlungen bei einer Bonusbank in die Formel (24) ergibt

$$\begin{aligned} EV^{HH,BB} = &(1-\tau_c)\left(x_2^H - \alpha w_1^{H,BB} - w_2^{H,BB}\right) - \\ &k\tau_c\left(\alpha w_1^{H,BB} + w_2^{H,BB}\right) + \left(1+i_{\tau_c}\right) \\ &\left((1-\tau_c)\left(x_1^H - (1-\alpha)w_1^{H,BB}\right) - \right. \\ &\left. k\tau_c(1-\alpha)w_1^{H,BB}\right) \end{aligned} \tag{33}$$

[202] Da die Perioden nicht mehr unabhängig voneinander sind, hat der Agent mehrere Möglichkeiten. Er kann in der ersten Periode weiterhin a_1^j wählen, aber seine Handlung der zweiten Periode auf die Realisation der ersten Periode bedingen. So könnte er z. B., wenn $\tilde{x}_1 = x_1^H$ eintritt, a_2^H wählen, bei $\tilde{x}_1 = x_1^L$ aber a_2^L. Dadurch ergeben sich acht mögliche Strategien des Agenten im Gegensatz zu den vier Strategien bei einer periodischen Vergütung.

[203] Vgl. Bierbaum (2002), S. 9f.

[204] Vgl. Janocha (2014), S. 93, Fußnote 64. Steuerliche Wirkungen können zwar theoretisch den Vorteil von $w_1^{L,BB} > 0$ aus Prinzipalsicht beeinflussen, haben aber auf das Endvermögen nie eine stärkere Wirkung als der rein negative Effekt der Entlohnungszahlung.

[205] Im Anhang wird gezeigt, dass die Teilnahmebedingung nicht bindet und die Renten gleich hoch sind.

$$EV^{HL,BB} = (1-\tau_c)\left(x_2^L - w_2^{L,BB}\right) - k\tau_c w_2^{L,BB} + \left(1+i_{\tau_c}\right)\left((1-\tau_c)\left(x_1^H - (1-\alpha)w_1^{H,BB}\right) - k\tau_c(1-\alpha)w_1^{H,BB}\right) \quad (34)$$

Weil im Zustand (x_1^H, x_2^L) die Entlohnungszahlung der ersten Periode bei einer Bonusbank mit $(1-\alpha)w_1^{H,BB} < w_1^H$ kleiner ist, gilt $EV^{HL,BB} > EV^{HL}\left(a^H\right)$. Andersherum ist die Entlohnungszahlung bei einer Bonusbank im Zustand (x_1^H, x_2^H) mit $w_1^{H,BB} > w_1^H$ insgesamt größer, sodass $EV^{HH,BB} < EV^{HH}\left(a^H\right)$ resultiert.

Die Gewichtung der zustandsabhängigen Bonusbankendvermögen mit den entsprechenden Wahrscheinlichkeiten und Einsetzen von $w_2^{L,BB} = 0$ ergibt[206]

$$EV^{BB}\left(a^H\right) = (1-\tau_d)\Big[(1-\tau_c)\left(\left(1+i_{\tau_c}\right)E(\tilde{x}_1) + E(\tilde{x}_2)\right) - \phi^H(1-\tau_c(1-k))\left(\left(\phi^H\alpha + (1-\alpha)\left(1+i_{\tau_c}\right)\right)w_1^{H,BB} + w_2^{H,BB}\right)\Big]. \quad (35)$$

Damit sind die relevanten Größen bestimmt, um die Vorteilhaftigkeit einer Bonusbank im Vergleich zu den Alternativen der periodischen Ausschüttung und des Grundgehalts zu untersuchen.

4.2.3. Vorteilhaftigkeitsvergleich und steuerinduzierte Wirkungen

Die Endvermögensdifferenz $EV^{BB}(a^H) - EV(a^L)$ gibt Aufschluss darüber, wann die Bonusbank einem Fixlohn aus Prinzipalsicht vorzuziehen ist. Dies ist der Fall, wenn[207]

$$\Delta\phi\left(\left(1+i_{\tau_c}\right)\Delta x_1 + \Delta x_2\right) \geq \phi^H\left(1 + \frac{k\tau_c}{1-\tau_c}\right) w_1^H\left(\frac{\phi^H\alpha + (1-\alpha)\left(1+i_{\tau_c}\right)}{1-\alpha+\delta\alpha\phi^H} + \frac{1}{\delta}\right). \quad (36)$$

Dem Vorteil des wahrscheinlicheren, hohen Cash-Flows in beiden Perioden (linke Seite) steht der Nachteil der Entlohnungszahlung im Erfolgsfall (rechte Seite) bei Gebrauch der Bonusbank gegenüber. Die Dividendensteuer τ_d mindert beide Seiten um denselben Faktor und wird bei der Ermittlung vernachlässigt. Die Bedingung (36) bestätigt die Erkenntnisse des einperiodigen Modells, die in der Bedingung (15) zum Ausdruck gebracht werden. Insbesondere wird auch die Bonusbankvergütung durch eine beschränkte Abzugsfähigkeit $k > 0$ mit $\tau_c > 0$ im Vergleich zum Grundgehalt benachteiligt, da die Entlohnungszahlungen größer sind. Zudem steigt w_1^H mit zunehmenden Einkommensteuersatz τ_{inc}, sodass eine erhöhte Steuerlast beim Manager zulasten des Prinzipals geht. Eine zusätzliche Belastung der Implementierung einer permanenten Anstrengung im Vergleich zu einem Grundgehalt wird durch $\delta^{-1} \geq 1$ deutlich.[208] Diese Zeitkomponente ergibt sich daraus, dass das Arbeitsleid v^H der zweiten Periode nicht mit δ diskontiert wird. Die Anreizsetzung für den Prinzipal wird dementsprechend teurer und der Prinzipal wählt die ex ante effiziente Strategie seltener.

Der Endvermögensvergleich zwischen der Bonusbank und der periodischen Vergütung ist mit Blick auf die in Kapitel 2.2.3 dargestellten Vorzüge der Bonusbank interessant. Die Bonusbank ist aus Eignersicht vorteilhaft, sofern

$$EV^{BB}(a^H) - EV(a^H) \geq 0 \iff \underbrace{(1-\tau_d)(\phi^H)^2\alpha w_1^{H,BB}(1-\tau_c(1-k))}_{\geq 0}\left(\underbrace{\delta(1+i(1-\tau_c))-1}_{\gtreqless 0}\right) \geq 0 \quad (37)$$

erfüllt ist.[209] Welches der beiden Anreizsysteme vorteilhaft ist, wird durch den Faktor $\delta(1+i(1-\tau_c))-1$ determiniert.[210] Sofern dieser vorzeichenbestimmende Faktor positiv ist, ist die Bonusbank vorzuziehen. Andernfalls erzielt der Prinzipal bei einer periodischen Vergütung ein höheres Endvermögen. In dem Fall $\delta = 1$ und $i = 0$ ist der Prinzipal indifferent zwischen der Nutzung einer Bonusbank und einer periodischen Vergütung, da die Bonusbank eine Zahlungsverschiebung ohne Vor- und Nachteil bewirkt. Steuerliche Wirkungen, die über die einperiodige Vertragsbeziehung hinausgehen, treten durch die Mehrperiodigkeit bei diesen Parameterwerten nicht auf.

In dem Vorzeichenfaktor wird der zentrale Trade-off des Grundmodells für den in der Arbeit zentralen Fall $\delta < 1$ und $i > 0$ ersichtlich. Die Bonusbank bewirkt auf der einen Seite eine verzögerte Auszahlung des Anteils α, die durch den höheren Kapitalmarktanlagebetrag zu einem zusätzlichen Zinsertrag $1+i(1-\tau_c)$ auf Unternehmensebene führt. Auf der anderen Seite muss bei einer Bonusbank das Entgelt der ersten Periode größer sein, damit entsprechende Leistungsanreize auch bei einem Manager mit kurzfristiger Konsumpräferenz δ gesetzt werden.

Der Unternehmensteuersatz τ_c beeinflusst hierbei den nachsteuerlich erwirtschafteten Zins direkt. Er benachteiligt die Bonusbank dadurch, dass der Zinsertrag einen steuerbaren Gewinn der Kapitalgesellschaft darstellt. In der Tabelle 4 werden die für den Prinzipal entscheidungsrelevanten Bruttozinssätze in Abhängigkeit vom Unternehmensteuersatz angegeben. Übersteigen die Zinssätze die in der Tabelle aufgeführten kritischen Werte, so ist die Bonusbank vorteil-

[206] Siehe Anhang.
[207] Siehe Anhang.
[208] Siehe dazu die Bedingung im Anhang.
[209] Siehe Anhang.
[210] Bierbaum (2002) untersucht in seinem zweiperiodigen Modell den Fall $\delta = 1$ und $i = 0$, ohne explizit eine Bonusbank zu untersuchen. Janocha (2014) nimmt anders als in dem Modell dieser Arbeit eine Zeitpräferenz (nicht: Konsumpräferenz) des Agenten sowie $i = 0$ an. Zudem vernachlässigen beide Beiträge Steuereinflüsse.

Hebesatz	300%	350%	400%	450%	500%
τ_c	26,33%	28,08%	29,83%	31,58%	33,33%
	Kritische Bruttozinssätze i_{krit}				
$\delta = 0{,}91$	13,42%	13,75%	14,09%	14,45%	14,83%
$\delta = 0{,}93$	10,22%	10,46%	10,73%	11,00%	11,29%
$\delta = 0{,}95$	7,14%	7,32%	7,50%	7,69%	7,89%
$\delta = 0{,}97$	4,20%	4,30%	4,41%	4,52%	4,64%
$\delta = 0{,}99$	1,37%	1,40%	1,44%	1,48%	1,51%

Tabelle 4: Kritische Bruttozinssätze in Abhängigkeit von der Konsumpräferenzrate und dem Gewerbesteuerhebesatz (Eigene Darstellung)

haft.[211] Dabei wird zugrunde gelegt, dass für inländische Kapitalgesellschaften eine differierende Steuerbelastung nur aus verschiedenen Hebesätzen resultiert. Die Gewerbesteuermesszahl und der effektive Körperschaftsteuersatz entsprechen in der Höhe derjenigen im deutschen Steuerrecht.
Es wird deutlich, dass die Bruttozinssätze zum Teil signifikant hoch sein müssen, damit sich eine Bonusbankvereinbarung für den Prinzipal lohnt. Ist für den Agenten bspw. ein Konsum in der ersten Periode 100 GE, in der zweiten aber nur noch 95 GE wert ($\delta = 0,95$), so muss das Unternehmen bei einem Hebesatz von 400% bereits eine Anlagemöglichkeit auf dem Kapitalmarkt oder unternehmensinterne Projekte mit einer Mindestrendite von 7,5% finden. Je höher τ_c (bzw. δ) ist, desto größer (bzw. kleiner) muss die erwartete Mindestrendite sein.

Neben den vorzeichenbestimmenden Größen in (37) gibt es weitere Größen, die verstärkend auf die Höhe der Endvermögensdifferenz wirken. Zum einen steigt die Endvermögensdifferenz im einbehaltenen Anteil der ersten Periode sowie in der Wahrscheinlichkeit, dass unterschiedliche Zahlungen in beiden Perioden resultieren: $\left(\phi^H\right)^2 \alpha w_1^{H,BB}$. Dies ist einleuchtend, weil gerade das Eintreten dieses Entlohnungsunterschieds die Endvermögensdifferenz ausmacht. Zum anderen mindert die Einkommensteuer auf den Ausschüttungsbetrag τ_d den Zielfunktionswert des Prinzipals weiterhin ohne eine der Alternativen zu bevorzugen.

Darüber hinaus sinkt die absolute Endvermögensdifferenz im Unternehmensteuersatz auch neben dem direkten Vorzeicheneffekt. Dies liegt daran, dass die Vergütungen nicht in vollem Umfang negativ in das Endvermögen des Prinzipals eingehen, da eine steuerliche Entlastung durch die (teilweise) Abzugsfähigkeit als Betriebsausgabe vorgesehen ist. Dadurch sinkt der Einfluss der Entlohnungszahlungen auf die absolute Endvermögensdifferenz. Im Umkehrschluss folgt, dass durch den beschränkt abzugsfähigen Anteil $k > 0$ eine höhere Endvermögensdifferenz induziert wird. Die Abbildung 6 veranschaulicht die Gesamtwirkung des Unternehmensteuersatzes auf die absolute Endvermögensdifferenz $EV^{BB}(a^H) - EV(a^H)$ für ausgewählte Parameterwerte k.
Zugrunde liegen die Parameterwerte aller Fälle aus der Tabelle 5 mit Ausnahme der variierten Werte τ_c und k. Die Differenz ist für $\tau_c \leq 0,3$ positiv, da der Zinsvorteil des in die Bonusbank eingestellten Anteils größer ist als der Nachteil durch die Konsumpräferenz des Agenten. Die unterschiedlichen Werte für k beeinflussen den Vorteil der Bonusbank in diesem Bereich nur marginal, weil die Wirkung der beschränkten Abzugsfähigkeit für kleine τ_c gering ist. Sofern die periodische Vergütung von Vorteil ist ($\tau_c > 0,3$), wird die Endvermögensdifferenz besonders für $k = 0,7$ immer geringer. Die steuerliche Entlastung durch die Absetzbarkeit als Betriebsausgabe wird durch die Belastung durch die beschränkte Abzugsfähigkeit für hohe τ_c dominiert, sodass sich die beschränkte Abzugsfähigkeit stärker auf den Vorteil einer periodischen Vergütung als auf den Vorteil einer Bonusbank auswirkt. Für $k = 0$ wird hingegen der Vorteil der periodischen Vergütung durch die Belastung des Nachsteuerzinssatzes für hohe τ_c sehr klein, zumal die dem Prinzipal zufließende Ausschüttung unbedeutend wird.

Insgesamt ist festzustellen, dass sowohl die Bonusbank und die periodische Vergütung als auch das Grundgehalt in einer langfristigen Vertragsbeziehung durch den Prinzipal genutzt werden. Dies hängt von den konkreten Parameterwerten ab. Die Tabelle 5 veranschaulicht dies, da in jeder der drei Parameterkonstellationen eine andere Vergütungsform durch den Prinzipal gewählt wird.[212]
In dem ersten Fall ist der Prinzipal indifferent zwischen der Nutzung der Bonusbank und der periodischen Vergütung.[213] Allerdings wird der Fixlohnvertrag durch den Prinzipal genutzt, da das Endvermögen mit $EV(a^L) = 105,79$ größer als die Endvermögen bei hoher induzierter Anstrengung ist. Im zweiten Fall gibt es zwei Änderungen zum vorherigen Fall, die in Fettdruck markiert sind. Zum einen wird unterstellt,

[211] Die Werte sind auf zwei Nachkommastellen gerundet. Der kritische Bruttozinssatz i_{krit} ergibt sich aus $\delta(1 + i(1-\tau_c)) - 1 \overset{!}{=} 0 \Rightarrow i_{krit} = \left(\delta^{-1} - 1\right)(1-\tau_c)^{-1}$.

[212] Sofern eine Variable mit dem Index t angegeben ist, gelten die Daten für $t = 1$ und $t = 2$.

[213] Deshalb gilt für $\tau_c = 0,3$ in der Abbildung 6: $EV^{BB}(a^H) - EV(a^H) = 0$.

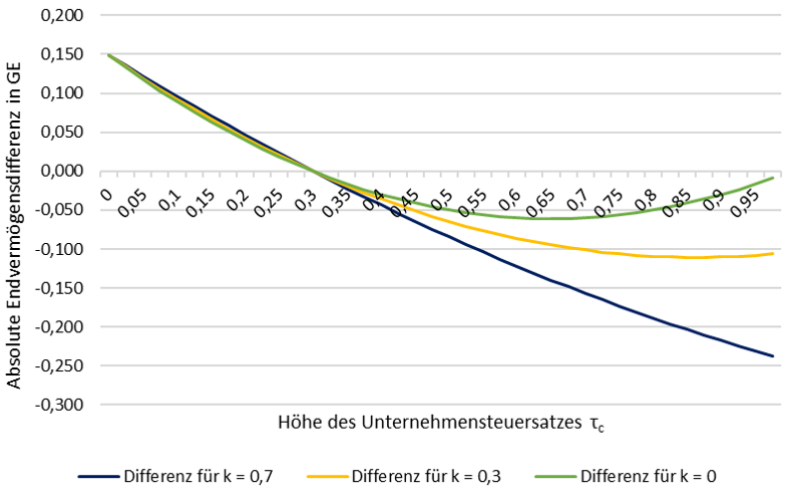


Abbildung 6: Endvermögensdifferenz $EV^{BB}(a^H) - EV(a^H)$ bei Variation des Unternehmensteuersatzes im Grundmodell (Eigene Darstellung)

Parameterkonstellationen				Zielgrößen			
	Fall 1	Fall 2	Fall 3		Fall 1	Fall 2	Fall 3
ϕ^H	0,75	0,75	0,75	F	0,00	0,00	0,00
ϕ^L	0,5	0,5	0,5	w_1^H	28,73	28,73	28,73
x_t^H	150	150	150	$w_1^{H,BB}$	33,55	33,55	33,55
x_t^L	50	50	50	$w_2^H = w_2^{H,BB}$	30,24	30,24	30,24
v^H	4	4	4	$w_t^L = w_t^{L,BB}$	0,00	0,00	0,00
α	0,5	0,5	0,5				
δ	0,95	0,95	0,95	$EV\left(a^L\right)$	**105,79**	121,34	90,34
i	0,0752	0,0752	0,0752	$EV\left(a^H\right)$	103,85	124,86	**92,96**
τ_{inc}	0,4431	0,4431	0,4431	$EV^{BB}\left(a^H\right)$	103,85	**124,90**	92,93
τ_c	0,3	**0,2**	**0,4**				
τ_d	0,26375	0,26375	0,26375				
k	0,5	**0**	0				

Zielgrößenwerte sowie i und τ_{inc} sind gerundet

Tabelle 5: Vorteilhaftigkeitsprüfung der verschiedenen Vergütungsstrukturen mit detaillierter Parameter- und Zielgrößenangabe (Eigene Darstellung)

dass nunmehr keine beschränkte Abzugsfähigkeit des Managergehalts vorliegt. Dies befördert unmittelbar die beiden Alternativen zur Implementierung der hohen Anstrengung. Die Senkung des Unternehmensteuersatzes auf $\tau_c = 20\%$ begünstigt die Bonusbank unmittelbar, da der nachsteuerliche Zinseffekt den Nachteil der verzögerten Auszahlung übersteigt. Die Bonusbank besitzt im zweiten Fall mit $EV^{BB}(a^H) = 124,90$ das höchste Endvermögen. Im dritten Fall erhöht sich im Vergleich zum zweiten Fall lediglich der Unternehmensteuersatz auf 40%, sodass der Nettozins der Kapitalgesellschaft nicht mehr ausreicht, um den Nachteil des bevorzugten Gegenwartskonsums zu übersteigen. Der Prinzipal maximiert sein Endvermögen durch die periodische Vergütung mit $EV(a^H) = 92,96$.

Der Vergleich der Endvermögenshöhen wurde als Instrument zur Identifizierung der Vorteilhaftigkeit einer Alternative genutzt. Schwieriger gestaltet sich jedoch die Beantwortung der Frage, wie hoch bspw. der Vorteil der Einführung einer Bonusbank in monetären Größen tatsächlich ist. Im zweiten Fall ist die naheliegende Antwort verglichen mit der periodischen Vergütung "0,04 GE", die naturgemäß davon abhängt, in welcher Größenordnung die Geldeinheiten angegeben werden. In einer großen Kapitalgesellschaft können Zinseffekte durchaus mit Geldbeträgen einhergehen, die trotz der eher geringen Differenzen in dem Zahlenbeispiel erhebliche Wirkungen entfalten. Dies gilt insbesondere, wenn die Bonusbank über einen längeren Zeitraum als die zwei Perioden im Modell genutzt wird. Problematisch in dem Zusammen-

hang ist, dass Prinzipal-Agenten-Modelle zwar Wirkrichtungen aufzeigen können, aber durch eine zum Teil restriktive Prämissensetzung eher weniger dafür geeignet sind, tatsächliche, monetäre Effekte anzugeben. Dies muss stets bei der Interpretation etwaiger Endvermögensdifferenzen berücksichtigt werden.

4.3. Modellerweiterung mit periodenabhängiger Einkommensteuer

4.3.1. Endvermögens- und Entlohnungsbestimmung

Im Grundmodell ist vereinfachend von einem konstanten Einkommensteuersatz τ_{inc} beim Manager in beiden Perioden ausgegangen worden. Dieser bildet den Einkommensteuertarif des § 32a EStG jedoch nur unzureichend ab. In Kapitel 2.3.1 wurde deutlich gemacht, dass die tariflichen Grenzeinkommensteuersätze mit zunehmendem zvE steigen. Diese Progressionswirkung beim Manager soll im Folgenden durch periodenabhängige Einkommensteuersätze $\tau_{inc,t}$ approximiert werden.[214] Gleichzeitig wird der direkte Einfluss von gesetzgeberischen Änderungen der tariflichen Einkommensteuer bestimmt. Zudem werden nachfolgend periodenabhängige Wahrscheinlichkeiten ϕ_t^j berücksichtigt.

Die Bestimmung der Entlohnungen in der Erweiterung wird an dieser Stelle mit Verweis auf die analoge Ermittlung im Grundmodell verkürzt dargestellt. Dabei wird die Erkenntnis genutzt, dass die periodische Vergütung für $\alpha = 0$ aus den bestimmten Größen der Bonusbank hervorgeht. Die intertemporale Partizipationsbedingung sowie die Anreizbedingung im Erfolgsfall der ersten Periode binden nicht und werden vernachlässigt. Weiterhin gilt für die Bonuszahlung bei niedrigem Cash-Flow $w_t^L = w_t^{L,BB} = 0$.[215] Bei einer Bonusbank ergibt sich die Entlohnungszahlung der ersten Periode bei hohem Cash-Flow aus der bindenden Anreizbedingung (siehe Anhang) als

$$w_1^{H,BB} = \frac{v^H}{\Delta\phi_1} \frac{1}{\left(1-\tau_{inc,1}\right)(1-\alpha)+\left(1-\tau_{inc,2}\right)\phi_2^H\delta\alpha} . \quad (38)$$

Der Prinzipal muss bei der Wahl von $w_1^{H,BB}$ demgemäß die Einkommensteuersätze beider Perioden einbeziehen. Der Einfluss der Steuersätze wird dabei mit α bzw. $1-\alpha$ gewichtet.

Die Entlohnungszahlung der zweiten Periode resultiert aus der bindenden Anreizbedingung der zweiten Periode (siehe Anhang) als

$$w_2^{H,BB} = w_2^H = \frac{v^H}{\Delta\phi_2} \frac{1}{1-\tau_{inc,2}} \frac{1}{\delta} . \quad (39)$$

Das Endvermögen des Prinzipals bei Nutzung einer Bonusbank lautet:[216]

$$\begin{aligned} EV^{BB}\left(a^H\right) = &(1-\tau_d)\Big[(1-\tau_c) \\ &\left(\left(1+i_{\tau_c}\right)E(\tilde{x}_1)+E(\tilde{x}_2)\right)- \\ &(1-\tau_c(1-k)) \\ &\left(\phi_1^H(1-\alpha)w_1^{H,BB}\left(1+i_{\tau_c}\right)+\right. \\ &\left.\phi_2^H\left(w_2^{H,BB}+\phi_1^H\alpha w_1^{H,BB}\right)\right)\Big] \end{aligned} \quad (40)$$

Durch Setzen von $\alpha = 0$ ergibt sich des Weiteren das Endvermögen bei periodischer Vergütung

$$\begin{aligned} EV\left(a^H\right) = (1-\tau_d)\Big[&(1-\tau_c)\left(\left(1+i_{\tau_c}\right)E(\tilde{x}_1)+E(\tilde{x}_2)\right) \\ &-(1-\tau_c(1-k)) \\ &\left(\phi_1^H w_1^H\left(1+i_{\tau_c}\right)+\phi_2^H w_2^H\right)\Big]. \end{aligned} \quad (41)$$

Das Endvermögen bei Nutzung eines Fixlohnvertrages ist weiterhin durch (28) bestimmt. Der einzige Unterschied ist hierbei, dass die Wahrscheinlichkeiten ϕ_t^L im Erwartungswert nicht konstant über beide Perioden sein müssen.

4.3.2. Vorteilhaftigkeitsvergleich und steuerinduzierte Wirkungen

Beim Vergleich der Bonusbank mit dem Fixlohnvertrag treten keine neuen Wirkungen auf, die über die Bedingung (36) aus dem Grundmodell hinausgehen. Die Vorteilhaftigkeitsbedingung wird durch die periodenabhängigen Einkommensteuersätze und Wahrscheinlichkeiten lediglich komplexer. Einkommensteuersatzerhöhungen jeder Periode benachteiligen die Anreizsysteme zur Implementierung eines hohen Anstrengungsniveaus weiterhin.

Steuerwirkungen der Erweiterung offenbaren sich hingegen in der Endvermögensdifferenz $EV^{BB}(a^H) - EV(a^H)$. Die Bonusbank ist gegenüber einer periodischen Vergütung vorteilhaft, wenn[217]

$$\delta\left(1+i_{\tau_c}\right)\left(1-\tau_{inc,2}\right)\left(1-\tau_{inc,1}\right)^{-1}-1 \geq 0. \quad (42)$$

Die Bedingung (42) zeigt, dass im Allgemeinen weiterhin keine uneingeschränkte Vorteilhaftigkeit einer der Alternativen besteht. Außerdem beeinflussen die periodenabhängigen Wahrscheinlichkeiten ϕ_t^H den Gebrauch der Bonusbank nicht. Allerdings beeinflussen nun die Einkommensteuersätze der Perioden die Vorteilhaftigkeit der Bonusbank bzw. der periodischen Vergütung unmittelbar. Während der Prinzipal im Grundmodell für $\delta\left(1+i_{\tau_c}\right) = 1$ indifferent zwischen den beiden Alternativen ist, ist die Bonusbank in dieser Konstellation bei periodenabhängiger Einkommensteuer für $\tau_{inc,1} \geq \tau_{inc,2}$ vorteilhaft.

Das Ergebnis entspricht dem intuitiv erwarteten Resultat. Die anteilige Einbehaltung von $w_1^{H,BB}$ bewirkt einen erhöhten, erwarteten Zufluss und damit eine erhöhte, erwartete Steuerzahlung beim Manager in $t = 2$. Demzufolge wird

[214] Aufgrund der Risikoneutralität von Prinzipal und Agent können die deterministischen Steuersätze $\tau_{inc,t}$ auch als erwartete Steuersätze interpretiert werden.

[215] Im Anhang werden die Anreizkompatibilitätsbedingungen hergeleitet.

[216] Siehe Anhang.

[217] Siehe Anhang.

$\tau_{inc,2}$ in der Bedingung (38) für die Festlegung von $w_1^{H,BB}$ anders als bei der periodischen Vergütung w_1^H berücksichtigt. Hintergrund des Einflusses von $\tau_{inc,t}$ auf die Zielfunktion des Prinzipals ist die bereits im einperiodigen Modell gereifte Erkenntnis, dass der Agent in einem Modell mit binärer Arbeitsanstrengung die Einkommensteuerbelastung an den Prinzipal überwälzt. Daher wählt der Prinzipal diejenige Alternative, bei der auch der Manager die geringste Einkommensteuerbelastung zu tragen hat.

Wenn durch den Gesetzgeber zukünftige Steuerentlastungen bspw. durch die Abschaffung des Solidaritätszuschlages in $t = 2$ durchgesetzt werden und der Prinzipal dies ebenfalls erwartet, so wählt er die Bonusbankvereinbarung häufiger. Andersherum sprechen in Zukunft erwartete Steuersatzerhöhung für die Nutzung einer periodischen Vergütung. Diese Überlegungen gelten indes nur für (erwartete) Änderungen innerhalb des Vereinbarungszeitraums, die zum Zeitpunkt des Vertragsschlusses feststehen.

Sofern $\delta\left(1+i_{\tau_c}\right) \neq 1$ gilt, muss der Prinzipal neben den periodischen Einkommensteuersätzen des Managers gleichzeitig den Trade-off des Grundmodells in seine Überlegungen einbeziehen. Es besteht sowohl die Möglichkeit, dass periodenabhängige Einkommensteuern die Vorteilhaftigkeit der Bonusbank oder der periodischen Vergütung intensivieren, als auch dass sie diese abschwächen oder gar zu einem Vorteilhaftigkeitswechsel führen. Die Abbildung 7 veranschaulicht beispielhaft den Einfluss der Steuersatzänderung $\tau_{inc,2}$ auf das Endvermögen aller Vertragsalternativen bei gegebenem Steuersatz $\tau_{inc,1}$. Dabei werden die über der Abbildung angegebenen Daten zugrunde gelegt.

Im Grundmodell ist die Bonusbank für die angegebene Parameterkonstellation bei einem Bruttozinssatz von knapp über 7,50% gegenüber einer sofortigen Bonusauszahlung im Vorteil (siehe Tabelle 4). Da in den Daten ein Bruttozins von 10% gegeben ist, lieferte die Bonusbank im Grundmodell stets ein höheres Endvermögen als die periodische Vergütung. Für Steuersätze $\tau_{inc,2} \in [0, 0.42)$ ist die Bonusbank die für den Prinzipal optimale Alternative. Gleichzeitig wird die Vorteilhaftigkeit der Bonusbank gegenüber der periodischen Vergütung verstärkt, weil die Bonusbank die Vergütungen in die steuerlich günstigere Periode verschiebt. Liegt $\tau_{inc,2}$ in dem Intervall $[0.42, 0.4294)$, so schwächt sich die Vorteilhaftigkeit der Bonusbank gegenüber der periodischen Vergütung ab, da der für die Bonusbank wichtigere Steuersatz den Vergleichssteuersatz $\tau_{inc,1}$ überschreitet. Für $\tau_{inc,2} \in [0.4294, 0.6638)$ ist die periodische Vergütung die bestmögliche Alternative.[218] Schließlich dominiert für sehr hohe $\tau_{inc,2} > 0.6638$ der Fixlohnvertrag die Alternativen zur Implementierung der hohen Anstrengung. Ursächlich dafür ist, dass der Prinzipal die hohe Einkommensteuerbelastung des Agenten trägt, während wegen $F = 0$ der Fixlohnvertrag unabhängig von jeglicher Einkommensteuer ist.

Das Zahlenbeispiel der Abbildung unterstreicht den erheblichen Effekt von Einkommensteuersatzänderungen. Der Vorteil, dass die Bruttoverzinsung im Grundmodell um etwa 2,5% höher liegt als der kritische Zinssatz i_{krit}, verschwindet bereits bei einer Erhöhung von $\tau_{inc,2}$ um ca. 0,94%. Die Tabelle 6 präzisiert die Änderungen von i_{krit} bei einer Erhöhung von $\tau_{inc,1} = 0,42$ auf $\tau_{inc,1} = 0,43$ im Vergleich zum Grundmodell.[219]

Die Erhöhung der kritischen Zinssätze schwankt zwischen 2,41% und 2,89% und steigt mit zunehmendem Unternehmensteuersatz und sinkender Konsumpräferenzrate. Während im Grundmodell bspw. für $\delta = 0,95$ und $\tau_c = 0,2983$ der kritische Zins i_{krit} bei 7,50% liegt, erhöht dieser sich durch den gestiegenen Steuersatz $\tau_{inc,2}$ um 2,63% auf 10,13%. Im umgekehrten Fall einer Steuersatzsenkung von $\tau_{inc,1} = 0,43$ auf $\tau_{inc,2} = 0,42$ sinken die kritischen Zinssätze zwischen 2,36% (für $\delta = 0,99$ und $\tau_c = 0,2633$) und 2,84%. Als Faustformel ist damit festzuhalten, dass eine Erhöhung bzw. Senkung der Steuer $\tau_{inc,1}$ um 1% mit einer Erhöhung bzw. Senkung des kritischen Zinssatzes um etwa 2,5% einhergeht.

4.3.3. Progressionseffekte

Neben der direkten Interpretation der gewählten Modellierung lassen periodenabhängige Einkommensteuersätze auch indirekt Rückschlüsse auf Progressionseffekte zu. Hintergrund der Untersuchung von Progressionseffekten ist die mögliche Zusammenballungswirkung des in die Bonusbank eingestellten Anteils $\alpha w_1^{H,BB}$. Ist dieser Betrag hoch, so kann bei Erfolg in $t = 2$ zusätzlich zu dem Bonus $w_2^{H,BB}$ ein deutlich höheres zvE anfallen als bei der periodischen Vergütung. Der beim Manager relevante Grenzeinkommensteuersatz steigt dann in $t = 2$. Umgekehrt besteht die Möglichkeit, dass durch die Einstellung von $\alpha w_1^{H,BB}$ in die Bonusbank der Grenzeinkommensteuersatz bei einer periodischen Vergütung in $t = 1$ höher ist als bei der Bonusbankvereinbarung.

Es zeigt sich, dass Progressionseffekte definitiv einen Einfluss auf die Vertragswahl des Prinzipals haben. Auf der einen Seite begünstigt die Bonusbank durch den einbehaltenen Anteil eine Glättung von Bonuszahlungen über den Vereinbarungszeitraum, sofern der Zustand (x_1^H, x_2^L) eintritt. Dieser Effekt wird in einem steigenden $\tau_{inc,1}$ eingefangen. Auf der anderen Seite kann im Zustand (x_1^H, x_2^H) die Bonusbank zu zusammengeballten Einkünften führen, die durch einen steigenden Steuersatz $\tau_{inc,2}$ abgebildet werden. Der Prinzipal sollte beide Effekte berücksichtigen und in sein Entscheidungskalkül einbetten.[220] Sämtliche durch die Mo-

[218] Diese Intervallgrenzen wurden auf vier Nachkommastellen gerundet.

[219] Die Werte in der Tabelle sind auf zwei Nachkommastellen gerundet worden. Der kritische Bruttozinssatz i_{krit} bei periodenabhängiger Einkommensteuer ergibt sich aus $\delta\left(1+i_{\tau_c}\right)\left(1-\tau_{inc,2}\right)-\left(1-\tau_{inc,1}\right) \overset{!}{=} 0 \Rightarrow i_{krit} = \left(\left(1-\tau_{inc,1}\right)\left(\delta\left(1-\tau_{inc,2}\right)\right)^{-1}-1\right)(1-\tau_c)^{-1}$.

[220] Alternativ könnte modelliert werden, dass eine Ausschüttung des Anteils $\alpha w_1^{H,BB}$ in der zweiten Periode einem separaten Einkommensteuersatz unterliegt, der größer ist als der Einkommensteuersatz auf die übrigen Einkünfte $(1-\alpha)w_1^{H,BB}, w_1^H$ und w_2^H. Dies ähnelte der Modellierung der Bonussteuer in Meißner et al. (2014). Der Nachteil dieser Modellierung wäre jedoch, dass Glättungseffekte der Bonusbank durch Steuersatzsenkungen in der ersten Periode nicht berücksichtigt würden.

x_t^H	x_t^L	ϕ_t^H	ϕ_t^L	v^H	α	δ	i	$\tau_{inc,1}$	τ_c	τ_d	k
150	50	0,65	0,4	4	0,5	0,95	0,1	0,42	0,3	0,25	0

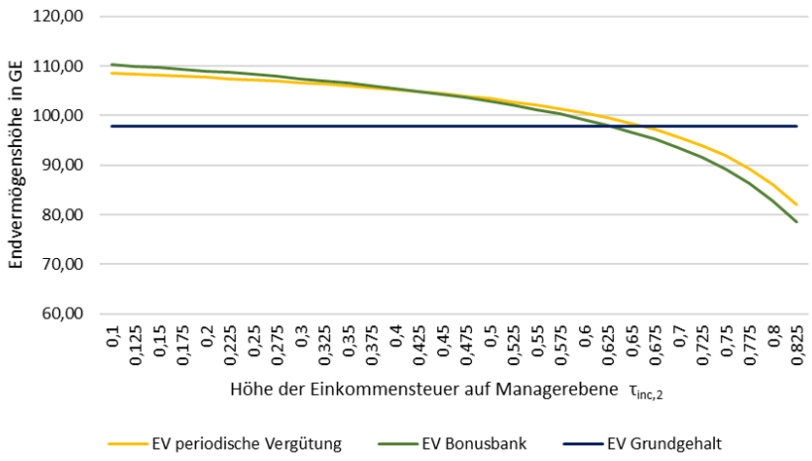


Abbildung 7: Endvermögenshöhen der Vertragsalternativen bei Variation des Einkommensteuersatzes $\tau_{inc,2}$ (Eigene Darstellung)

Hebesatz	300%	350%	400%	450%	500%
τ_c	26,33%	28,08%	29,83%	31,58%	33,33%
	Änderungen von i_{krit}				
$\delta = 0{,}91$	+2,62%	+2,68%	+2,75%	+2,82%	+2,89%
$\delta = 0{,}93$	+2,56%	+2,62%	+2,69%	+2,76%	+2,83%
$\delta = 0{,}95$	+2,51%	+2,57%	+2,63%	+2,70%	+2,77%
$\delta = 0{,}97$	+2,45%	+2,51%	+2,58%	+2,64%	+2,71%
$\delta = 0{,}99$	+2,41%	+2,46%	+2,53%	+2,59%	+2,66%

Tabelle 6: Änderungen der kritischen Bruttozinssätze in Abhängigkeit von der Konsumpräferenzrate und dem Gewerbesteuerhebesatz für $\tau_{inc,1} = 0,42$ und $\tau_{inc,2} = 0,43$ (Eigene Darstellung)

dellierung angedeuteten Progressionseffekte in diesem Abschnitt sollten allerdings als Maximalwirkungen angesehen werden, da Steuersatzerhöhungen das gesamte Einkommen der Periode betreffen, während die Progression tatsächlich nur Auswirkungen auf Entlohnungsanteile hat.

Der Gesamteffekt der Progression ist in dem gewählten Modellrahmen nur schwer quantifizierbar. Zum einen liegt dies daran, dass die Progression über perioden- und nicht über bemessungsgrundlagenabhängige Steuersätze abgebildet wird. Zum anderen hängen Progressionseffekte von der übrigen Einkommenssituation des Managers in den jeweiligen Perioden ab, die im Rahmen des Partialkalküls ausgeblendet wird. Weder ist sichergestellt, dass der Manager durch den einbehaltenen Teil $\alpha w_1^{H,BB}$ bei einer Bonusbank einen niedrigeren Grenzsteuersatz in $t = 1$ abzuführen hat, noch dass er bei Erfolg in $t = 2$ einen höheren Grenzsteuersatz zahlen muss.

Es lässt sich jedoch insgesamt die Tendenz aussprechen, dass die Bonusbank eher einen Progressionsvorteil als einen Progressionsnachteil induziert. Bonusbanken werden in der Praxis über einen längeren Zeitraum als die hier modellierten zwei Perioden vereinbart. Wird in jeder der Perioden des (längeren) Zeitraums ein konstanter Anteil α in die Bonusbank eingestellt, so bewirkt dies einen Glättungseffekt über alle Kooperationsperioden. Eine Ausnahme besteht lediglich, wenn der Manager in jeder Periode sein vorgegebenes Ziel erreicht und das restliche Guthaben in der letzten Periode ausgeschüttet wird. In diesem Fall entsteht in der letzten Periode ein Progressionsnachteil, der zwar bis zu einem maximalen Grenzsteuersatz von 45% zzgl. Solidaritätszuschlag

gedeckelt, aber gegenüber einer periodischen Ausschüttung nach den Erkenntnissen aus dem Modell unvorteilhaft ist.

4.4. Einbeziehung von Verlusten auf Unternehmensebene

4.4.1. Vorbemerkungen

In der bisherigen Analyse wurden Verluste auf Unternehmensebene stets durch eine geeignete Parameterwahl ausgeschlossen. Die steuerlich differenzierte Behandlung von Gewinnen und Verlusten ist allerdings ein wesentlicher Bestandteil des deutschen Steuersystems. Zugleich wird die Bonusbank als Instrument zur Verlustbeteiligung des Managers propagiert, sodass von Interesse ist, inwieweit defizitäre Situationen tatsächlich die Wahl für oder gegen eine Bonusbank beeinflussen.

In der Prinzipal-Agenten-Literatur zu Anreizwirkungen der Besteuerung wurde die asymmetrische Behandlung von Gewinnen und Verlusten lediglich in dem Aufsatz von Niemann (2011) untersucht. Die differenzierte Behandlung von steuerlichen Gewinnen und Verlusten führt in seinem einperiodigen Modell dazu, dass die leistungsabhängige Entlohnung zur Implementierung der hohen Anstrengung im Vergleich zu einem Grundgehalt begünstigt wird.[221] Kritisch kann aber die Modellierung der Verluste gesehen werden. Zeiteffekte, die durch Verlustverrechnungsbeschränkungen entstehen, werden in seinem Modell nicht explizit abgebildet. Dies ermöglicht hingegen das hiesige Mehrperiodenmodell.

Um eine sinnvolle Handhabbarkeit der Verlustvorschriften des § 10d EStG und des § 10a GewStG zu gewährleisten, werden in dieser Arbeit nicht alle Elemente dieser Vorschriften in den Modellrahmen integriert. Zum einen wird von der Möglichkeit eines körperschaftsteuerlichen Verlustrücktrages abgesehen. Dies erscheint nicht allzu restriktiv, da in großen Kapitalgesellschaften dieser Rücktragshöchstbetrag ohnehin eine untergeordnete Rolle spielt. Empirisch ist dies an den hohen Verlustvortragsbeständen zu erkennen.[222] Dadurch entfällt die differenzierte Behandlung von körperschaftsteuerlichen und gewerbesteuerlichen Verlusten, die durch den unerlaubten Rücktrag von Gewerbeverlusten im deutschen Steuerrecht eigentlich vorgesehen ist. Zum anderen wird grundsätzlich von einem unbeschränkten Verlustvortrag ausgegangen. Damit wird die Mindestbesteuerungsquote in Höhe von 40% des € 1.000.000 übersteigenden Betrages nicht explizit modelliert. Diese Annahme wäre unproblematisch, wenn der Verlustvortragsbetrag der betrachteten Kapitalgesellschaft unter dem Sockelbetrag liegen würde. In großen Kapitalgesellschaft kann davon aber nicht grundsätzlich ausgegangen werden. Die Arbeit verliert dadurch bei der Abbildung des deutschen Steuerrechts an Genauigkeit.

Ein weiteres Problem, welches unter Umständen eintritt, ist die Bewertung von steuerlichen Verlustvorträgen am Ende des Kooperationszeitraums. Steuerliche Verluste könnten in $t = 1$ so hoch sein, dass selbst Gewinne in $t = 2$ diese nicht ausgleichen können. Bei einer Verlustentstehung in $t = 2$ sind zudem verbleibende Verlustvorträge aufgrund der zweiperiodigen Vertragsbeziehung und des ausgeschlossenen Verlustrücktrages garantiert. Zur Bewertung dieser steuerlichen Verlustvorträge wird eine künstliche, dritte Periode eingeführt. Das Unternehmen existiert nach der Vertragsbeziehung weiterhin und erwirtschaftet in $t = 3$ einen deterministischen Cash-Flow $x_3 > 0$, der zum Ausgleich möglicher Verlustvorträge aus vorherigen Perioden besteht. Dieser Cash-Flow ist unabhängig von der Leistung eines Agenten und dient ausschließlich als Hilfskonstrukt.

Um die Künstlichkeit des Cash-Flows x_3 abzuschwächen, wird der Parameter $\theta \in (0, 1]$ eingeführt. Dieser Parameter bildet die zukünftige Ertragslage des Unternehmens ab. Für $\theta = 1$ ist der Cash-Flow x_3 so groß, dass sämtliche verbleibende Verlustvorträge in $t = 3$ ausgeglichen werden. Ist $\theta < 1$, so können nicht alle Verlustvorträge in $t = 3$ ausgeglichen werden. Diese müssten über die dritte Periode hinaus vorgetragen werden. Zur Approximation der damit einhergehenden Zinsverluste erfolgt daher eine Steuererstattung in Höhe von $\theta\tau_c$ je Geldeinheit des Verlustvortrages in $t = 3$. θ reflektiert folglich eine Quote der zukünftigen Verlustabzugsfähigkeit bezogen auf den Zeitpunkt $t = 3$.[223] Da $x_3 > 0$ aber unabhängig von der durch den Prinzipal gewählten Vertragsform anfällt, ist dieser Cash-Flow nicht entscheidungsrelevant. Daher wird er nur indirekt über den Parameter θ in der Endvermögensermittlung berücksichtigt.

Die zeitliche Abfolge aus dem Mehrperiodenfall in der Abbildung 5 ändert sich durch die Einbeziehung von Verlusten insoweit, dass sämtliche Überschüsse aus dem Unternehmen erst zum Ende der dritten Periode an den Prinzipal ausgeschüttet werden. Angehäufte Überschüsse am Ende von $t = 2$ werden am Kapitalmarkt angelegt und stellen steuerbare Zinseinkünfte in $t = 3$ dar. Die ermittelten Endvermögenshöhen werden damit nun auf den Bewertungszeitpunkt $T = 3$ bezogen. Zur Vereinfachung wird wieder von $\phi_t^H = \phi^H$, $\phi_t^L = \phi^L$ und $\tau_{inc,t} = \tau_{inc}$ analog zum Grundmodell ausgegangen. Weil durch steuerliche Verluste auf Unternehmensebene die Nebenbedingungen des Vertragsproblems nicht tangiert werden, sind die Entlohnungszahlungen des Grundmodells ebenfalls unter Einbeziehung von Verlusten optimal.

Bei der Untersuchung des Einflusses von Verlustregelungen wird als Vergleichsmaßstab die symmetrische Besteuerung von Verlusten zugrunde gelegt. Eine symmetrische Behandlung von steuerlichen Verlusten läge vor, wenn das Unternehmen bei Verlusten in der selben Periode eine Steuererstattung erhalten würde und keine zeitverzögerte Erstattung, wie sie durch Verlustvorträge bewirkt wird. Der Fall der symmetrischen Besteuerung ist in den Endvermögenshöhen des Grundmodells abgebildet, wenn negative Cash-Flows möglich sind. Daher können Steuerwirkungen durch Verluste aus dem Vergleich mit den Endvermögenshöhen des Grundmodells (symmetrische Besteuerung) extrahiert werden.

[221] Vgl. Niemann (2011), S. 13f.
[222] Vgl. Niemann (2004), S. 360.
[223] Vgl. ähnlich in Mehrmann und Sureth-Sloane (2017), S. 7; Niemann (2004), S. 368.

4.4.2. Verlustentstehung in der ersten Periode

Relevante Konstellationen und Endvermögensbestimmung. Zunächst wird der Fall untersucht, in dem in der ersten Periode ein Verlust auftreten kann, während in der zweiten Periode die Verlustentstehung weiterhin ausgeschlossen ist. Steuerliche Verluste in $t = 1$ werden in diesem Modell über negative Cash-Flows $x_1^L < 0$ bei niedrigem Output abgebildet.

Nach Niemann (2011) sind andere Konstellationen, in denen Verluste auftreten, ökonomisch uninteressant. Zum einen könnten sowohl bei niedrigem als auch bei hohem Cash-Flow in einer Periode Verluste auftreten. Dann wäre aber die Unterlassungsalternative für die betrachtete Periode optimal. Zum anderen wäre eine Verlustentstehung ausschließlich bei hohem Output denkbar. Hierbei ist bekannt, dass die Bemessungsgrundlage des Unternehmens in $t = 1$ bei einer periodischen Vergütung durch $x_1^H - (1-k)w_1^H$ gegeben ist.[224] Wenn dagegen ein steuerlicher Verlust bei hohem Cash-Flow $x_1^H - (1-k)w_1^H < 0$ und ein Gewinn bei niedrigem Cash-Flow $x_1^L \geq 0$ gleichzeitig erfüllt wären, würde nach der Erkenntnis aus dem Anhang die Implementierung einer hohen Anstrengung immer durch die niedrige Anstrengung dominiert.[225] Damit ist der Fall $x_1^L < 0$ der einzig relevante Fall in dieser Arbeit.

Durch die Zahlungsabflüsse bei niedrigem Output in $t = 1$ können steuerliche Verluste nicht unabhängig davon untersucht werden, aus welchen Mitteln die Zahlungsabflüsse gedeckt werden. Es wird davon ausgegangen, dass diese Zahlungsabflüsse durch liquide Mittel $L > 0$ aus anderen Unternehmensbereichen beglichen werden. Diese liegen am Ende von $t = 1$ vor. Zur Vereinfachung gelte $L = -x_1^L$. Sofern sich in $t = 1$ ein hoher Output realisiert, sind die Überschüsse $x_1^H - w_1^H$ positiv, sodass die liquiden Mittel am Kapitalmarkt angelegt werden. Die dominierte Unterlassungsalternative liefert dem Prinzipal nun ein Endvermögen von $EV = (1 + i_{\tau_c})^2 L$.

Der Vorteilhaftigkeitsvergleich zwischen der Bonusbank und der periodischen Vergütung erübrigt sich bei ausschließlicher Verlustmöglichkeit in $t = 1$. Der Grund dafür liegt darin, dass die zustandsabhängigen Endvermögen $EV^{Lm,BB}(a^H)$ und EV^{Lm} mit $m \in \{L, H\}$, die durch die steuerlichen Verlustverrechnungsbeschränkungen betroffen sind, ohnehin aufgrund gleicher Entlohnungszahlungen identisch sind. Hingegen bleiben die für die Endvermögensdifferenz relevanten Zustände $EV^{Hm,BB}$ und $EV^{Hm}(a^H)$ durch die asymmetrische Besteuerung unberührt. Seien $EV_{T=2}^{BB}(a^H)$ und $EV_{T=2}(a^H)$ die Endvermögen aus dem Grundmodell und $EV_{T=3}^{BB}(a^H)$ und $EV_{T=3}(a^H)$ diejenigen Endvermögen unter Einbeziehung von Verlusten.[226] Dann lässt sich die Erkenntnis formal folgendermaßen ausdrücken:

$$\begin{aligned}&\left(1 + i_{\tau_c}\right)\left(EV_{T=2}^{BB}\left(a^H\right) - EV_{T=2}\left(a^H\right)\right)\\ =&EV_{T=3}^{BB}\left(a^H\right) - EV_{T=3}\left(a^H\right) \ .\end{aligned} \tag{43}$$

Der Faktor $(1 + i_{\tau_c})$ auf der linken Seite von (43) führt zu einer einheitlichen Bewertung der Endvermögensdifferenzen auf den Zeitpunkt $T = 3$. Es wird deutlich, dass durch $x_1^L < 0$ keine steuerliche Wirkung auftritt, die über den Trade-off des Grundmodells hinausgeht. Zwar sinken die Endvermögenshöhen im Vergleich zum Fall der symmetrischen Besteuerung, aber dies bei der Bonusbank und bei der periodischen Vergütung im gleichen Maße. Der Prinzipal muss weiterhin den Vorteil der Verzinsung durch die Einbehaltung von $\alpha w_1^{H,BB}$ gegen die Präferenz des Managers, früh konsumieren zu wollen, abwägen (Bedingung (37)).

Steuerliche Effekte ergeben sich aus der Gegenüberstellung der Bonusbank und dem Fixlohnvertrag, sodass die zustandsabhängigen Endvermögen nun detaillierter betrachtet werden. Da Verluste in diesem Abschnitt lediglich durch $x_1^L < 0$ auftreten können, bleiben die zustandsabhängigen Endvermögen $EV^{Hm,BB}$ und $EV^{Hm}\left(a^L\right)$ durch die Verlustverrechnungsbeschränkungen unberührt. Der einzige Unterschied zum Grundmodell besteht darin, dass die liquiden Mittel im Endvermögen berücksichtigt werden müssen und eine Verzinsung der Überschüsse von $t = 2$ nach $t = 3$ durch die Kapitalmarktanlage erfolgt. Das angepasste Endvermögen vor Ausschüttung an den Eigner entspricht zum Bewertungszeitpunkt $T = 3$:

$$EV_{T=3}^{Hm,BB} = \left(1 + i_{\tau_c}\right) EV_{T=2}^{Hm,BB} + \left(1 + i_{\tau_c}\right)^2 L \ . \tag{44}$$

Wird $EV^{Hm,BB}$ in (44) durch $EV^{Hm}\left(a^L\right)$ ersetzt, so gilt die Bedingung analog bei Zahlung eines Grundgehalts.

Die zustandsabhängigen Endvermögenshöhen $EV_{T=3}^{Lm,BB}$ und $EV_{T=3}^{Lm}\left(a^L\right)$ müssen über Finanzpläne bestimmt werden. In der Tabelle 7 wird der Finanzplan bei leistungsabhängiger Vergütung ohne die dritte Periode aufgezeigt, wenn in $t = 1$ ein Verlust vorliegt.[227] Ziel dieses unvollständigen Finanzplans ist es, alle denkbaren Fallkombinationen hinsichtlich des Unternehmensgewinns bzw. -verlusts in $t = 2$ zu präsentieren, die daraufhin in der Tabelle 8 zusammengefasst werden.

Durch den Finanzplan wird ersichtlich, dass die Bestimmung der Endvermögen von der Parameterkonstellation und dem realisierten Zustand in $t = 2$ abhängt. Im Allgemeinen kann das zvE in $t = 2$ (Position [5]) positiv oder negativ sein. Ist es positiv, so wird eine Steuerzahlung fällig und es verbleiben keine Verlustvorträge. Der KM-Anlagebetrag ist in dem Fall durch $(1 - \tau_c)(x_2^m - w_2^m) - k\tau_c w_2^m - \tau_c x_1^L$ gegeben.[229]

[224] Sie ist wegen $(1-\alpha)w_1^{H,BB} < w_1^H$ kleiner als die Bemessungsgrundlage $x_1^H - (1-k)(1-\alpha)w_1^{H,BB}$ bei einer Bonusbank.

[225] Vgl. Niemann (2011), S. 13. Für eine Bonusbank gilt die Erkenntnis ebenfalls, da die Ausgangssituation mit $x_1^H - (1-k)(1-\alpha)w_1^{H,BB} < 0$ noch restriktiver sein müsste.

[226] Im Vergleich zum Grundmodell wird die Notation hier durch Einfügen des Bewertungszeitpunkts T im Index angepasst, um den Unterschied zur asymmetrischen Besteuerung bei Verlustmöglichkeit zu verdeutlichen.

[227] Die Endvermögensbestimmung der Bonusbank ist äquivalent zur Endvermögensbestimmung der periodischen Vergütung, da die Entgelte im Zustand (x_1^L, x_2^m) identisch sind.

[228] Da die Fallbezeichnung in Kapitel 4.4.3 ähnlich ist, steht die "I" für die mögliche Verlustentstehung in der ersten Periode.

[229] Ähnlich wie im Finanzplan des Abschnitts 4.2.1 ergibt sich der Anlagebetrag aus [10]=[1]-[7]-[8]+[9].

Periode		t=1	t=2
Ermittlung der Steuerlast			
[1]	Cash-Flow	x_1^L	x_2^m
[2]	Entlohnung anteilig	0	$-(1-k)\,w_2^m$
[3]	KM-Zinsen	0	0
[4]	Verlustvortrag Vorjahr	0	$+x_1^L$
[5]	Gewerbeertrag = zvE	$= x_1^L < 0$	$= x_2^m - (1-k)\,w_2^m + x_1^L \gtreqless 0$
[6]	Verlustvortrag	$-x_1^L$	$\max\{0; -\left(x_2^m - (1-k)\,w_2^m + x_1^L\right)\}$
[7]	Unternehmensteuern	0	$\max\{0; \tau_c[x_2^m - (1-k)\,w_2^m + x_1^L]\}$
Ermittlung der Überschüsse			
[8]	Entlohnung	$w_1^L = 0$	w_2^m
[9]	Habenzins und Tilgung	0	0
[10]	KM-Anlagebetrag	$L + x_1^L = 0$	$\min\{x_2^m - w_2^m;$ $(1-\tau_c)\left(x_2^m - w_2^m\right) - k\tau_c w_2^m - \tau_c x_1^L\}$
[11]	Saldo/Endvermögen	0	0

Tabelle 7: Ausschnitt des Finanzplans im Zustand (x_1^L, x_2^m) bei Verlust in t=1 und leistungsabhängiger Vergütung (Eigene Darstellung)

Fall-bezeichnung[228]	(x_1^L, x_2^H) und GG	(x_1^L, x_2^H) und BB/PV	(x_1^L, x_2^L) und BB/PV/GG	Bedingung
Gewinnfall I	Gewinn	Gewinn	Gewinn	$x_2^L \geq -x_1^L$
Hybridfall I	Gewinn	Verlust	Verlust	$x_2^H \geq -x_1^L$ und $x_2^H - (1-k)w_2^H < -x_1^L$
Verlustfall I	*Verlust*	*Verlust*	*Verlust*	$x_2^H < -x_1^L$

Tabelle 8: Fälle bei möglicher Verlustentstehung in $t = 1$ bezogen auf das zvE in $t = 2$ (Eigene Darstellung)

Ist das zvE negativ, so wird der Betrag in die nächste Periode vorgetragen und es wird keine Steuerzahlung fällig. Am Kapitalmarkt können dann $x_2^m - w_2^m$ GE angelegt werden. Bei ausschließlicher Betrachtung der leistungsabhängigen Vergütung kann in $t = 2$ folglich ein Verlustübertrag verbleiben oder nicht.

Interessant ist nun, wann sich in $t = 2$ die steuerlichen Konsequenzen im Zustand (x_1^L, x_2^m) bei der Endvermögensermittlung des Grundgehalts im Vergleich zum leistungsabhängigen Entgelt unterscheiden.[230] Gleiche steuerliche Konsequenzen liegen vor, sofern der Zustand (x_1^L, x_2^L) eintritt. In diesem Zustand stimmen die Unternehmensgewinne (Position [5]) in der Periode $t = 2$ wegen $w_t^{L,BB} = w_t^L = F = 0$ überein. Somit gilt auch stets $EV_{T=3}^{LL,BB} = EV_{T=3}^{LL}\left(a^H\right) = EV_{T=3}^{LL}\left(a^L\right)$. Im Zustand (x_1^L, x_2^H) sind die Endvermögenshöhen nicht identisch, weil bei der leistungsabhängigen Vergütung das Entgelt $w_2^H > F = 0$ gezahlt wird. Die Vergütung selbst führt unter Umständen dazu, dass in $t = 2$ nicht ausgeglichene Verlustvorträge verbleiben.

Die Tabelle 8 fasst sämtliche Fallkombinationen hinsichtlich der steuerlichen Konsequenzen in der zweiten Periode in Abhängigkeit von der gewählten Vergütung (BB, PV und GG) zusammen.

Im Gewinnfall I der zweiten Periode bei realisiertem $x_1^L < 0$ wird der Verlustvortrag der ersten Periode unabhängig vom realisierten Output der zweiten Periode ausgeglichen. Wenn ein positives zvE im Zustand (x_1^L, x_2^L) gesichert ist, so ist dies auch stets für (x_1^L, x_2^H) der Fall. Daher ist die relevante Bedingung $x_2^L \geq -x_1^L$. Im Verlustfall I liegt unabhängig von der Entgeltpolitik ein steuerlicher Verlust in $t = 2$ vor. $x_2^H < -x_1^L$ stellt sicher, dass ein Verlust beim Grundgehalt im Zustand

[230] Unter identischen, steuerlichen Konsequenzen wird hier das gleichzeitige Vorliegen eines (bzw. keines) Verlustübertrags in $t = 2$ bei der Bonusbank und dem Fixlohnvertrag verstanden. Unterschiedliche, steuerliche Konsequenzen liegen demgegenüber vor, wenn ein Verlustübertrag bei der Bonusbank, nicht aber beim Fixlohnvertrag verbleibt.

(x_1^L, x_2^H) und damit automatisch in allen anderen Fällen vorliegt. Der auffälligste Fall ist der Hybridfall I. Er zeichnet sich dadurch aus, dass wegen der Entlohnungszahlung zur Implementierung der hohen Anstrengung ein vollständiger Verlustausgleich nur beim Grundgehalt möglich ist.

Inwieweit sich die Vorteilhaftigkeit einer Bonusbank im Vergleich zu einem Grundgehalt durch Verlustverrechnungsbeschränkungen ändert, wird anhand des Gewinnfalls I und des Hybridfalls I untersucht. Der Gewinnfall I hat verglichen mit dem Fall der symmetrischen Besteuerung die kleinsten Unterschiede aufzuweisen, hebt jedoch bereits die zentrale Steuerwirkung heraus. Der Hybridfall I ist auffällig, weil er gerade den Unterschied zwischen der Bonusbank und dem Grundgehalt unterstreicht. Intuitiv wäre zu erwarten, dass durch die steuerlich asymmetrische Behandlung von Gewinnen und Verlusten leistungsabhängige Vergütungen im Hybridfall I benachteiligt werden. Dies gilt es zu prüfen.

Steuerinduzierte Wirkungen im Gewinnfall I.
Im Gewinnfall I der zweiten Periode lauten die Endvermögen für $\tilde{x}_1 = x_1^L$ bei einer Bonusbank bzw. dem Grundgehalt vor Ausschüttung an den Eigner

$$EV_{T=3}^{Lm,BB} = \left(1+i_{\tau_c}\right)\left((1-\tau_c)\left(x_2^m - w_2^m\right) - k\tau_c w_2^m - \tau_c x_1^L\right) \quad (45)$$

$$EV_{T=3}^{Lm}\left(a^L\right) = \left(1+i_{\tau_c}\right)\left((1-\tau_c)x_2^m - \tau_c x_1^L\right) . \quad (46)$$

Aus den Gleichungen (45) und (46) wird deutlich, dass der einzige Unterschied zum Fall der symmetrischen Besteuerung im Gewinnfall I darin liegt, dass die Steuererstattung $-\tau_c x_1^L$ in der zweiten und nicht in der ersten Periode erfolgt.[231]

Nach Gewichtung der zustandsabhängigen Endvermögen mit den entsprechenden Wahrscheinlichkeiten ergibt sich das Endvermögen bei Nutzung der Bonusbank im Gewinnfall I für $\tau_d = 0$ durch

$$\begin{aligned} EV_{T=3}^{BB}\left(a^H\right) = \left(1+i_{\tau_c}\right)\Big[& (1-\tau_c)\left(\left(1+i_{\tau_c}\right)\phi^H x_1^H + E(\tilde{x}_2)\right) - x_1^L\left(\phi^H\left(1+i_{\tau_c}\right) + \left(1-\phi^H\right)\tau_c\right) \\ & - \phi^H(1-\tau_c(1-k)) \cdot \\ & \left(\left(\phi^H\alpha + (1-\alpha)\left(1+i_{\tau_c}\right)\right)w_1^{H,BB} + w_2^{H,BB}\right)\Big] . \end{aligned} \quad (47)$$

Für das Grundgehalt ergibt sich

$$\begin{aligned} EV_{T=3}\left(a^L\right) = \left(1+i_{\tau_c}\right)\Big[& (1-\tau_c)\left(\left(1+i_{\tau_c}\right)\phi^L x_1^H + E(\tilde{x}_2)\right) - x_1^L\left(\phi^L\left(1+i_{\tau_c}\right) + \left(1-\phi^L\right)\tau_c\right)\Big] . \end{aligned} \quad (48)$$

Durch den vorgesehenen Verlustvortrag sinken bei Verlusteintritt die Endvermögen beim Prinzipal für alle Vergütungssysteme im Vergleich zum Grundmodell, da die verzögerte Steuererstattung einen Zinsverlust auf den Betrag $-\tau_c x_1^L$ bewirkt. Formal bedeutet dies am Beispiel der Bonusbank:

$$EV_{T=3}^{BB}\left(a^H\right) < \left(1+i_{\tau_c}\right)\left(EV_{T=2}^{BB}\left(a^H\right) + \left(1+i_{\tau_c}\right)L\right) . \quad (49)$$

Der Zinsnachteil tritt abhängig von der induzierten Anstrengung mit Wahrscheinlichkeit $\left(1-\phi^j\right)$ auf, während bei hohem Cash-Flow in $t = 1$ die liquiden Mittel $L = -x_1^L$ am Kapitalmarkt angelegt werden. Dies wird ebenfalls in den erwarteten Endvermögenshöhen (47) und (48) deutlich.

Von entscheidender Bedeutung ist, welche der Alternativen hierdurch stärker benachteiligt wird. Die Bonusbank ist im Gewinnfall I für $EV_{T=3}^{BB}\left(a^H\right) \geq EV_{T=3}\left(a^L\right)$ vorteilhaft. Die Vorteilhaftigkeitsbedingung lautet

$$\begin{aligned} & \Delta\phi\left((1-\tau_c)\left(\Delta x_2 + \left(1+i_{\tau_c}\right)x_1^H - \left(1+i_{\tau_c}-\tau_c\right)x_1^L\right)\right) \geq \\ & \phi^H(1-\tau_c(1-k))\,w_1^H\left(\frac{\phi^H\alpha + (1-\alpha)\left(1+i_{\tau_c}\right)}{1-\alpha+\delta\alpha\phi^H} + \frac{1}{\delta}\right). \end{aligned} \quad (50)$$

Die Bedingung (50) ähnelt der Bedingung (36) aus dem Grundmodell. Insbesondere ist der Entlohnungsterm der rechten Seite in beiden Bedingungen identisch. Lediglich die linken Seiten der Vorteilhaftigkeitsbedingungen unterscheiden sich. Im Anhang wird gezeigt, dass die linke Seite der Bedingung (50) größer ist als diejenige der Bedingung (36). Die Vorteilhaftigkeitsbedingung im Gewinnfall I ist folglich weniger einschränkend als im Grundmodell und die Bonusbank wird für $i \neq 0$ und $\tau_c \neq 0$ häufiger als im Fall der symmetrischen Besteuerung genutzt. Somit begünstigen Verlustverrechnungsbeschränkungen die Nutzung der Bonusbank.

Das Ergebnis kann dadurch erklärt werden, dass bei der Nutzung des Grundgehalts der Verlustfall $x_1^L < 0$ häufiger auftritt als bei der Nutzung der Bonusbank. Dies spiegelt sich in der positiven Wahrscheinlichkeitsdifferenz $\Delta\phi$ wider. Da im deutschen Steuerrecht keine sofortige Steuererstattung vorgesehen ist, entgeht dem Prinzipal bei der Nutzung eines fixen Grundgehalts im Vergleich zur symmetrischen Besteuerung der Zins auf den Erstattungsbetrag für $i > 0$ und $\tau_c > 0$ häufiger. Offensichtlich haben Verlustvorträge keine Auswirkung auf den Zielfunktionswert des Prinzipals, sofern $i = 0$ ist. In diesem Fall hat die zeitverzögerte Erstattung keine Auswirkung, weil keine Opportunitätskosten durch eine Alternativanlage vorliegen. Außerdem gibt es im Gewinnfall I keine erhöhte Nutzung der Bonusbank durch den Prinzipal, sofern $\tau_c = 0$ bzw. $\tau_c \to 1$ erfüllt ist. Dies ist einleuchtend, zumal für $\tau_c = 0$ keine Steuererstattung aus $x_1^L < 0$ resultiert und $\tau_c \to 1$ den nachsteuerlichen Zins extrem klein werden lässt.

Im Vergleich zum Fall der symmetrischen Besteuerung können signifikante Änderungen durch den Zinseffekt erzeugt werden, wenn die Endvermögensdifferenz zwischen der Bonusbank und dem Grundgehalt sehr klein ist. Insofern kann die größere Wahrscheinlichkeit eines Zinsverlustes bei dem Grundgehalt ausschlaggebend für die Nutzung der Bonusbank sein. Die Abbildung 8 veranschaulicht dies beispielhaft mit den nachfolgend angegebenen Daten.

[231] Der daraus resultierende Zinsnachteil ist ein direkter Effekt der ausschließlichen Möglichkeit eines Verlustvortrages. Vgl. grundlegend Schneider (1977), S. 643.

x_t^H	x_1^L	x_2^L	ϕ_t^H	ϕ_t^L	v^H	α	δ	i	$\tau_{inc,t}$	τ_d	k
150	-100	100	0,65	0,45	5,4	0,5	0,95	0,1	0,42	0,25	0

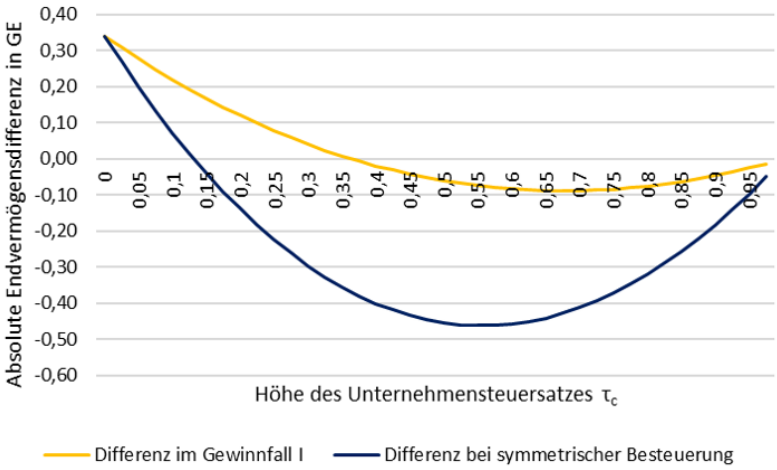


Abbildung 8: Endvermögensdifferenzen $EV^{BB}(a^H) - EV(a^L)$ bei symmetrischer und asymmetrischer Besteuerung (Gewinnfall I) in Abhängigkeit vom Unternehmensteuersatz (Eigene Darstellung)

Ein höherer Unternehmensteuersatz hat bei symmetrischer Besteuerung für $k = 0$ zwei Effekte auf die Endvermögensdifferenz. Zum einen sinkt der Zinsertrag auf den wahrscheinlicheren, hohen Cash-Flow der ersten Periode, was sich in dem fallenden Teil der Kurve zeigt. Zum anderen wird im steigenden Teil der Kurve deutlich, dass sich der Zinsnachteil einer Entlohnungszahlung in $t = 1$ für hohe τ_c reduziert.[232] Für die angegebenen Daten ist die Bonusbank bei symmetrischer Besteuerung für $\tau_c \in [0, 0.1315)$ optimal. Ab dem Steuersatz $\tau_c \geq 0.1315$ wird jedoch das Fixgehalt vorteilhaft, da der Nachteil des wahrscheinlicheren, hohen Cash-Flows dominiert. Der absolute Endvermögensnachteil der Bonusbank bleibt erhalten, schmilzt aber für hohe τ_c wegen des dominierenden Entlohnungseffektes ab.
Die asymmetrische Verlustbehandlung steigert für alle $\tau_c \in (0, 1)$ die Vorteilhaftigkeit der Bonusbank. Deshalb liegt die Kurve im Gewinnfall I stets über derjenigen bei symmetrischer Besteuerung. Ein Steuersatz von $\tau_c > 0,5$ steigert zwar den Steuererstattungsbetrag $-\tau_c x_1^L$, senkt aber den erwirtschafteten Nettozins i_{τ_c}. Für $\tau_c < 0,5$ ist das genaue Gegenteil der Fall. Durch die Komplementarität dieser Effekte liegt der maximale Vorteil der Bonusbank durch die verzögerte Steuererstattung daher bei einem Steuersatz von $\tau_c = 0,5$ vor.

Kein Unternehmen eines Prinzipals mit Sitz in Deutschland würde bei symmetrischer Besteuerung in dieser Datenkonstellation die Bonusbank nutzen, da alleine der Körperschaftsteuersatz die kritischen 13,15% übersteigt. Hingegen ist die Bonusbank im Gewinnfall I für alle Steuersätze $\tau_c \in [0, 0.3644)$ vorteilhaft. Die Verlustverrechnungsbeschränkungen führen dazu, dass jedes Unternehmen mit einem Gewerbesteuerhebesatz kleiner als 589% die Bonusbank bei asymmetrischer Behandlung verwendet.[233] Folglich beeinflusst der Zinseffekt nun auch die Vorteilhaftigkeit im Vergleich zum Grundgehalt.

Änderungen der anderen steuerlichen (und nichtsteuerlichen) Parameter bleiben von ihrer Wirkrichtung im Gewinnfall I identisch zum Grundmodell. Die absolute Höhe des Zinsvorteils der asymmetrischen Besteuerung ändert sich bei Variation von k und τ_{inc} nicht. Allerdings zeigt sich, dass der Einfluss der beschränkten Abzugsfähigkeit $k > 0$ in Verbindung mit $\tau_c > 0$ stets eine stärkere Auswirkung hat als der Einfluss von τ_c auf die Zinseffekte bei symmetrischer Besteuerung. Dies wird in der Abbildung 9 veranschaulicht.
Unter leicht angepasster Datenkonstellation mit $x_t^H = 160$ und $k = 0,1$ wird dies dadurch ersichtlich, dass die Endvermögensdifferenzen fortan monoton fallende Funktionen im Unternehmensteuersatz sind. Bei symmetrischer Besteuerung ist der Fixlohnvertrag für $\tau_c > 0.3502$ optimal, bei asymmetrischer Besteuerung für $\tau_c > 0.3906$. Die Differenz der kritischen Steuersätze geht auf den Zinsvorteil der Bonusbank aufgrund der Verlustmöglichkeit zurück, wird allerdings für steigende $k > 0$ stetig geringer als in der Abbildung 8.

Steuerinduzierte Wirkungen im Hybridfall I.
Im Gewinnfall I verbleiben weder bei der Fixlohnalternati-

[232] Die Effekte spiegeln sich in der Endvermögensdifferenz des Grundmodells in $\Delta\phi\left(1 + i_{\tau_c}\right)\Delta x_1$ bzw. $-\phi^H w_1^{H,BB}\left(\alpha\phi^H + (1-\alpha)\left(1 + i_{\tau_c}\right)\right)$ wider.

[233] Der Hebesatz ergibt sich implizit aus: $0,3644 \stackrel{!}{=} 0,15 \cdot 1,055 + 0,035 \cdot Hebesatz$.

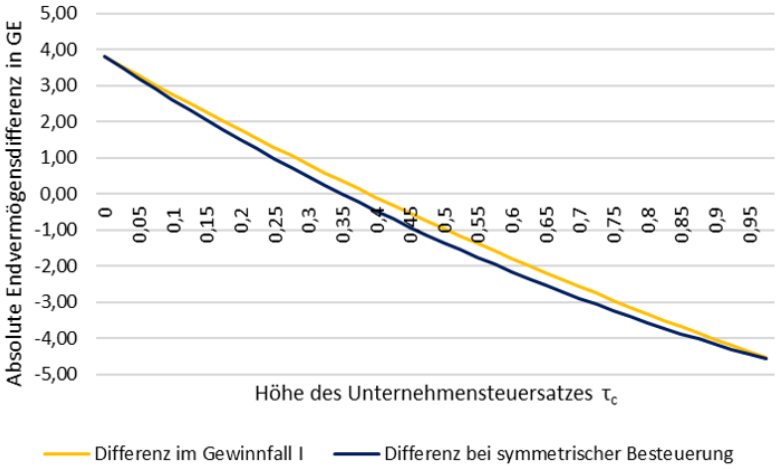


Abbildung 9: Endvermögensdifferenzen $EV^{BB}\left(a^H\right) - EV(a^L)$ bei symmetrischer und asymmetrischer Besteuerung (Gewinnfall I) unter angepasster Datenkonstellation (Eigene Darstellung)

ve noch bei der Bonusbankvereinbarung Verlustvorträge in $t = 2$, sofern ein Verlust $x_1^L < 0$ eintritt. Der Hybridfall I ist hingegen dadurch gekennzeichnet, dass im Zustand (x_1^L, x_2^L) bei beiden Vergütungsformen und im Zustand (x_1^L, x_2^H) lediglich bei der Bonusbank Vortragsüberträge in $t = 2$ verbleiben. Diese können erst in $t = 3$ ausgeglichen werden. Somit wird die künstliche dritte Periode anders als im Gewinnfall I als Hilfskonstrukt benötigt. Es wurde bereits festgestellt, dass die steuerlich asymmetrische Verlustbehandlung das Endvermögen des Prinzipals auch bei leistungsabhängigen Vergütungen mindert. Durch die Notwendigkeit des zusätzlichen Verlustvortrages in $t = 2$ im Zustand (x_1^L, x_2^H) ist das Ausmaß dieses Nachteils zu prüfen.

Der Verlustausgleich in der dritten Periode hängt von der zukünftigen Ertragslage $\theta \in (0, 1]$ ab. Wenn der verbleibende Verlustvortrag in $t = 2$ bspw. 100 GE beträgt, so werden in der dritten Periode Steuern in Höhe von $\tau_c \theta \cdot 100$ GE erstattet. Bei einer guten, zukünftigen Ertragslage $\theta = 1$ erfolgt in $t = 3$ eine komplette Steuererstattung, während sie für $\theta < 1$ lediglich anteilig ist. Zinserträge, die durch die Kapitalmarktanlage von $t = 2$ bis $t = 3$ angelegt werden, sind steuerbare Einkünfte und unterliegen in $t = 3$ analog dem Steuersatz $\tau_c \theta$.

Unverändert ergeben sich die zustandsabhängigen Endvermögen $EV_{T=3}^{Hm,BB}$ bzw. $EV_{T=3}^{Hm}\left(a^L\right)$ durch Gleichung (44). Das Endvermögen $EV_{T=3}^{Lm,BB}$ lässt sich über den Finanzplan in der Tabelle 9 ermitteln.
Vor Ausschüttung an den Prinzipal ergibt sich das zustandsabhängige Endvermögen bei der Bonusbankvereinbarung als:

$$EV_{T=3}^{Lm,BB} = (1 + i(1 - \tau_c \theta))\left(x_2^m - w_2^m\right) - \tau_c \theta \left(x_2^m - (1-k) w_2^m + x_1^L\right) . \quad (51)$$

Das Endvermögen bei Implementierung der niedrigen Anstrengung ist im Zustand (x_1^L, x_2^H) weiterhin durch Gleichung (46) aus dem Gewinnfall I gegeben. Lediglich bei Realisation des niedrigen Outputs in beiden Perioden ändert es sich zu

$$EV_{T=3}^{LL}\left(a^L\right) = (1 + i(1 - \tau_c \theta)) x_2^L - \tau_c \theta \left(x_2^L + x_1^L\right) . \quad (52)$$

Nachfolgend wird auf die explizite Berechnung der erwarteten Endvermögenshöhe in Abhängigkeit von der Vergütungsform und der Endvermögensdifferenz $EV_{T=3}^{BB} - EV_{T=3}\left(a^L\right)$ verzichtet. Eine kompakte Darstellung dieser Ausdrücke ist nicht möglich, sodass stattdessen Steuerwirkungen im Hybridfall I durch eine numerische Analyse hervorgehoben werden.

Wie im Gewinnfall I bleibt die Vorteilhaftigkeitsbedingung der Bonusbank verglichen mit der periodischen Vergütung identisch, weil beide Vergütungsformen durch die Verlustverrechnungsbeschränkungen in gleichem Umfang benachteiligt werden. Offensichtlich vermindern sich die Endvermögenshöhen aber in stärkerem Maße als im Gewinnfall I, da Zinsverluste durch Verlustvorträge am Ende von $t = 1$ und von $t = 2$ entstehen. Diese Zinsverluste sind für $\tau_c = 0,5$ maximal, sofern sämtliche Verluste in $t = 3$ ausgeglichen werden.[234] Voraussetzung hierfür ist eine ausreichend gute Ertragslage $\theta = 1$. Gehen verbleibende Verlustvortragsbeträge unter ($\theta < 1$), so entsteht dadurch eine zusätzliche Zielfunktionswertminderung beim Prinzipal, die mit zunehmendem Steuersatz τ_c steigt. Bei der Analyse der Bonusbank verglichen mit dem Grundgehalt werden diese Effekte näher quantifiziert.

Die zentrale Erkenntnis aus dem Hybridfall I ist die Tatsache, dass die Bonusbanknutzung gegenüber dem Fixlohnvertrag durch Verlustverrechnungsbeschränkungen begünstigt wird. Die gegen die Nutzung des Grundgehalts sprechende, höhere Eintrittswahrscheinlichkeit von Verlusten dominiert damit den potenziellen Nachteil der Bonusbank durch den

[234] Siehe dazu die Erklärung im Abschnitt 4.4.2.

	Periode	t=1	t=2	t=3
Ermittlung der Steuerlast				
[1]	Cash-Flow	x_1^L	x_2^m	$(x_3 > 0)$
[2]	Entlohnung anteilig	0	$-(1-k)w_2^m$	-
[3]	KM-Zinsen	0	0	$i(x_2^m - w_2^m)$
[4]	Verlustvortrag Vorjahr	0	$+x_1^L$	$x_2^m - (1-k)w_2^m + x_1^L$
[5]	Gewerbeertrag = zvE	$= x_1^L < 0$	$= x_2^m - (1-k)w_2^m$ $+x_1^L < 0$	$i(x_2^m - w_2^m) + x_2^m$ $-(1-k)w_2^m + x_1^L$
[6]	Verlustvortrag	$-x_1^L$	$-[x_2^m - (1-k)w_2^m + x_1^L]$	-
[7]	Unternehmensteuer	0	0	$\tau_c\theta[i(x_2^m - w_2^m)+$ $x_2^m - (1-k)w_2^m + x_1^L]$
Ermittlung der Überschüsse				
[8]	Entlohnung	0	w_2^m	-
[9]	Habenzins und Tilgung	0	0	$(1+i)(x_2^m - w_2^m)$
[10]	KM-Anlagebetrag	$L + x_1^L = 0$	$x_2^m - w_2^m$	-
[11]	Saldo/Endvermögen	0	0	$EV_{T=3}^{Lm}\left(a^H\right)$

Tabelle 9: Finanzplan im Zustand (x_1^L, x_2^m) im Hybridfall I bei Implementierung der permanenten Anstrengung in Abhängigkeit von der Ertragslage θ (Eigene Darstellung)

verbleibenden Verlustvortrag im Zustand (x_1^L, x_2^H) des Hybridfalls I. Um die Wirkungen explizit herauszuheben, werden die effektiven Endvermögensänderungen bei Realisation von $\tilde{x}_1 = x_1^L$ im Vergleich zur symmetrischen Besteuerung dargestellt. Für den Zustand (x_1^L, x_2^H) ist die effektive Endvermögensänderung bei der Bonusbank durch

$$\left(1-\phi^H\right)\phi^H(1-\tau_d)\left(\left(1+i_{\tau_c}\right)EV_{T=2}^{LH,BB} - EV_{T=3}^{LH,BB}\right) \quad (53)$$

gegeben. Beim Fixlohnvertrag und im Zustand (x_1^L, x_2^L) müssen die Wahrscheinlichkeiten und Endvermögen in (53) entsprechend ersetzt werden.

Die Abbildung 10 illustriert die effektive Endvermögensänderung im Zustand (x_1^L, x_2^H) für zwei ausgewählte Werte θ im Vergleich zur symmetrischen Besteuerung. Zugrunde liegen die in der Tabelle angegeben Parameterwerte, wobei Änderungen zum Gewinnfall I hervorgehoben sind.
Die effektiven Endvermögensdifferenzen sind stets positiv, da Verlustverrechnungsbeschränkungen das Endvermögen des Prinzipals mindern. Das zustandsabhängige Endvermögen beim Fixlohnvertrag ist unabhängig von θ, da in $t = 2$ keine verbleibenden Verlustvorträge vorliegen. Die Entlohnungszahlung w_2^H bei der Bonusbankvereinbarung führt dazu, dass in $t = 2$ Vortragsüberträge verbleiben, die erst in $t = 3$ ausgeglichen werden können. Zusammen mit dem negativen Zinseffekt durch den Vortrag von $t = 1$ nach $t = 2$ führt die asymmetrische Besteuerung zu einer effektiven Benachteiligung der Bonusbank, selbst wenn sämtliche Verlustvorträge unmittelbar in $t = 3$ ausgeglichen werden ($\theta = 1$). Daher liegt die blaue Kurve oberhalb der gelben Kurve. Allerdings muss festgestellt werden, dass der maximale Unterschied für $\tau_c = 0,5$ mit 0,09 GE vergleichsweise gering ist. Der monetäre Wert dieses Zinseffektes hängt jedoch von der Dimension ab, in der die Geldeinheiten gemessen werden und sollte daher trotzdem nicht vernachlässigt werden.

Beim Vergleich der effektiven Benachteiligungen für $\theta = 0,5$ ist der Nachteil der Bonusbank (grüne Kurve) deutlich stärker zu erkennen. Dies resultiert aus der schlechten, zukünftigen Ertragslage, die weitere Verlustvorträge außerhalb der drei betrachteten Perioden zur Folge hat. Generell ist $\theta < 1$ eine zusätzliche Restriktion der Verlustverrechnung. Die im deutschen Steuerrecht festgelegte Mindestbesteuerung kann bspw. in den Parameter θ einfließen, weil diese ebenfalls zu weiteren Zinsverlusten aufgrund der weiter in die Zukunft verschobenen Steuererstattung führt.[235] Unter Zugrundelegung eines (Top-)Managers, dessen Entlohnung w_2^H sehr hoch ist, können die resultierenden Negativzinseffekte signifikant sein.

Indes kann auch eine effektive Benachteiligung des Fixlohnvertrages im Zustand (x_1^L, x_2^H) gegeben sein. Wenn ceteris paribus bspw. $\phi^H > 0,7$ (bzw. $\phi^H > 0,768$) gilt, so ist der effektive Nachteil des Grundgehalts für alle τ_c und $\theta = 1$ (bzw. $\theta = 0,5$) größer, zumal bei der Bonusbank Verluste

[235] Die Modellierung bildet die Mindestbesteuerung des deutschen Steuerrechts nicht genau ab, da die Berücksichtigung der Mindestbesteuerung auch in der zweiten Periode hätte erfolgen müssen. Die Wirkung der Mindestbesteuerung wird gleichwohl abgebildet.

x_1^H	x_2^H	x_1^L	x_2^L	ϕ_t^H	ϕ_t^L	v^H	α	δ	i	$\tau_{inc,t}$	τ_d	k
200	**125**	-100	**0**	0,65	0,45	**6**	0,5	0,95	0,1	0,42	0,25	0

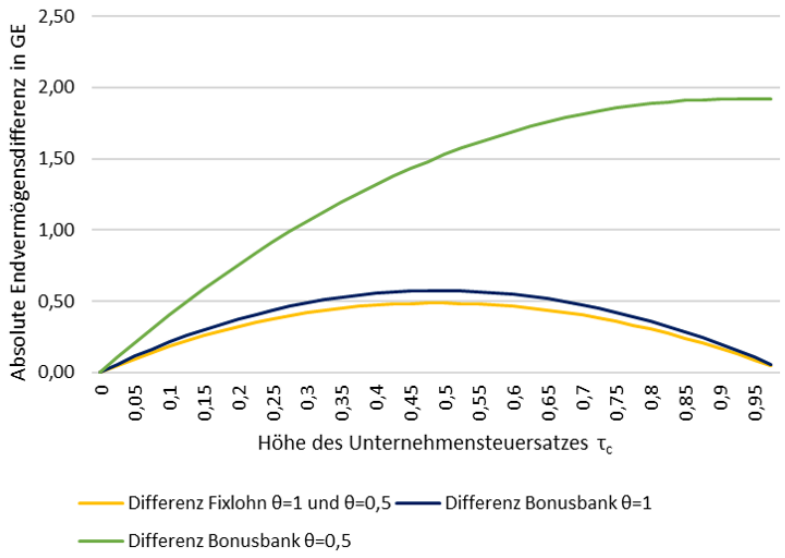


Abbildung 10: Effektive Endvermögensbenachteiligung durch die asymmetrische Besteuerung (Hybridfall I) im Zustand (x_1^L, x_2^H) (Eigene Darstellung)

seltener eintreten. Das Ergebnis der Benachteiligung des Fixlohnvertrages im Hybridfall I wäre für die angepassten Wahrscheinlichkeiten bestätigt, weil im Gewinnfall I eine Benachteiligung im Zustand (x_1^L, x_2^L) bereits identifiziert wurde. Zur Ergebnisherleitung wird aber im Folgenden von den in der Abbildung 10 zugrunde liegenden Daten ausgegangen, die aus Sicht der Bonusbanknutzung das "pessimistische" Szenario repräsentiert.

Je höher die tarifliche Einkommensteuer in diesem Szenario ist ($\tau_{inc} \uparrow$), umso stärker reduziert sich im Hybridfall I $EV_{T=3}^{LH,BB}$, da die Entlohnungszahlung und folglich die Verluste in $t = 2$ steigen. Dadurch wird die effektive Endvermögensbenachteiligung der Bonusbank im Zustand (x_1^L, x_2^H) verstärkt. Ist ein höherer Anteil der Entlohnung nicht als Betriebsausgabe abzugsfähig ($k \uparrow$), so sinkt der Nachteil der Verlustverrechnungsbeschränkungen bei Bonusbanknutzung in diesem Zustand. Dies liegt daran, dass der vorgetragene Verlust aus $t = 2$ geringer ist als bei einer vollständigen Abzugsfähigkeit. Während der Endvermögensnachteil der asymmetrischen Besteuerung $\left(\left(1+i_{\tau_c}\right)EV_{T=2}^{LH,BB} - EV_{T=3}^{LH,BB}\right)$ bspw. für $k = 0$ und $\tau_c = 0,3$ bei 0,49 GE liegt, so sinkt er für $k = 0,5$ auf 0,39 GE. Der effektive Vorteil beträgt damit 0,1 GE. Dieser führt jedoch nicht dazu, dass eine Erhöhung der beschränkten Abzugsfähigkeit das insgesamt erwartete Endvermögen des Prinzipals erhöht. Der bereits bekannte Nachteil der beschränkten Abzugsfähigkeit in den Zuständen (x_1^H, x_2^m) bewirkt stets einen stärkeren Nachteil hinsichtlich des erwarteten Endvermögens $EV_{T=3}^{BB}\left(a^H\right)$ als der induzierte, effektive Vorteil auf das zustandsabhängige Endvermögen $EV_{T=3}^{LH,BB}$.

Falls sich in beiden Perioden die niedrigen Outputniveaus realisieren, so sind die Endvermögenshöhen weiterhin gleich hoch: $EV_{T=3}^{LL,BB} = EV_{T=3}^{LL}\left(a^L\right)$. Dabei sinken die Endvermögenshöhen auf ein niedrigeres Niveau als im Gewinnfall I, da Verluste für $\theta = 1$ ebenfalls in die dritte Periode vorgetragen werden. Zusätzliche Restriktionen des Verlustausgleichs in der dritten Periode ($\theta \downarrow$) vermindern das Endvermögen weiter. Weil aber im Fixlohnvertrag der Fall (x_1^L, x_2^L) wahrscheinlicher ist, ist die effektive Endvermögenssenkung höher als bei der Bonusbank.

Für die Gesamtbeurteilung des Verlusteinflusses im Hybridfall I müssen die effektiven Endvermögenssenkungen beider Zustände zusammen betrachtet werden. In der Abbildung 11 werden diese kumulierten Endvermögenssenkungen bei der Bonusbank und dem Fixgehalt abgebildet.

Der Fixlohnvertrag wird für die obig gewählte Parameterkonstellation unabhängig von θ benachteiligt. Dies ist in der Abbildung daran zu erkennen, dass die Endvermögenssenkungen beim Fixlohnvertrag für die jeweils gleichen Parameterwerte θ für alle Unternehmensteuersätze größer sind als bei einer Bonusbank. Damit ändern sich die Ergebnisse im Vergleich zum Gewinnfall I insgesamt nicht. Der zentrale Grund hierfür ist, dass die Eintrittswahrscheinlichkeit für $\tilde{x}_1 = x_1^L$ wegen $\phi^L > \phi^H$ bei Implementierung der niedrigen Anstrengung größer ist. Der Nachteil der wahrscheinlicheren Verlustentstehung im Fixlohnvertrag ist größer als der Nach-

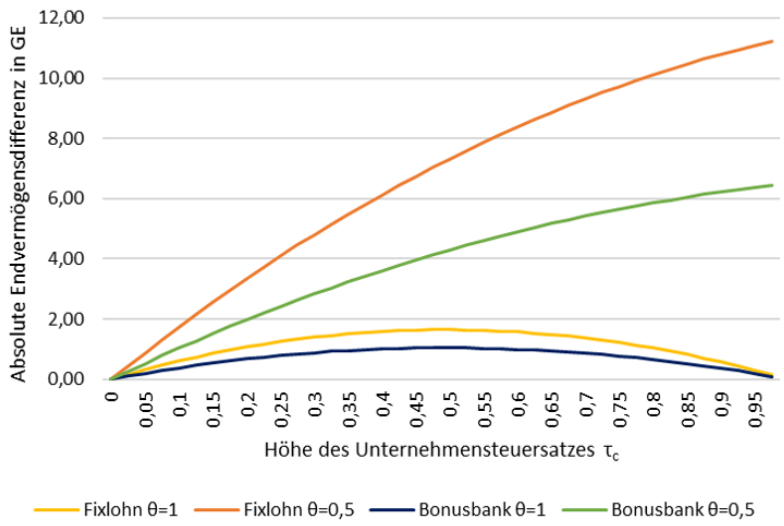


Abbildung 11: Gesamte Endvermögensbenachteiligung durch die asymmetrische Besteuerung (Hybridfall I) (Eigene Darstellung)

teil der Bonusbank im Zustand (x_1^L, x_2^H). Unabhängig davon, ob $\theta = 1$ oder $\theta = 0,5$ ist, gibt es keine zulässige Parameterkonstellation, in der sich diese Erkenntnis umkehren könnte.

Wenn bspw. in der dritten Periode alle Verluste ausgeglichen ($\theta = 1$) werden, so ist die Bonusbank für $\tau_{inc} = 0,85$ oder $v^H = 19$ effektiv durch die Verluste in $t = 1$ benachteiligt, weil das Entgelt w_2^H sehr hoch ist und daraus der negative Effekt im Zustand (x_1^L, x_2^H) signifikant wird. Für höhere Steuersätze als $\tau_{inc} = 0,5989$ beim Manager präferiert der Prinzipal aber in $t = 2$ immer den Fixlohnvertrag, da die Rückflüsse aus einer Bonusbankvereinbarung mit $x_2^H - \alpha w_1^{H,BB} - w_2^{H,BB} < x_2^L$ gegeben sind. Insbesondere würde der Prinzipal deshalb in $t = 2$ das hohe Anstrengungsniveau bei der Bonusbankvereinbarung nicht mehr implementieren. Dies steht allerdings im Widerspruch zu der Annahme in der Arbeit, dass der Prinzipal durch die leistungsabhängigen Vergütungen eine permanente Anstrengung des Agenten erreichen möchte.[236] Der selbe Widerspruch lässt sich für den Fall $\theta = 0,5$ herleiten.

Insgesamt wird damit bei einer Verlustentstehung in der ersten Periode deutlich, dass aus Verlustverrechnungsbeschränkungen entstehende Zeiteffekte bei der Bewertung von Vergütungssystemen nicht vernachlässigt werden sollten. Die Vorteilhaftigkeit des Fixlohnvertrags als Alternative zur Bonusbank sinkt, während der Vergleich von Bonusbank und periodischer Vergütung nicht beeinflusst wird. Im Verlustfall I (siehe Tabelle 8) sind keine divergierenden Erkenntnisse zu erwarten, da der Hybridfall I bereits den "maximalen" Unterschied der Bonusbank zum Fixlohnvertrag heraushebt. Die Benachteiligung des Fixlohnvertrages gilt analog im Vergleich zur periodischen Vergütung. Damit wird das Ergebnis aus Niemanns einperiodigem Modell, in dem er Zeiteffekte durch Bemessungsgrundlageneffekte approximiert, bestätigt. Er kommt ebenfalls zu dem Schluss, dass die differenzierte Behandlung von Gewinnen und Verlusten leistungsabhängige Verträge fördert.[237]

4.4.3. Verlustentstehung in der zweiten Periode

Relevante Konstellationen und Endvermögensbestimmung. Bisher wurde eine Verlustentstehung in $t = 2$ ausgeschlossen. Steuerliche Verluste in $t = 2$ haben bisher nur aufgrund von nicht ausgeglichenen Verlustvorträgen der ersten Periode vorgelegen. Um den Einfluss der Verlustentstehung in $t = 2$ von der Verlustentstehung in $t = 1$ zu separieren, wird im Folgenden wieder von $x_1^L \geq 0$ ausgegangen. Als Ursache für die Verlustentstehung in $t = 2$ wird ähnlich wie im Abschnitt 4.4.2 ein negativer Cash-Flow $x_2^L < 0$ unterstellt. Weiterhin werden Zahlungsabflüsse durch am Ende der ersten Periode vorliegende liquide Mittel gedeckt.[238]

Eine Vorteilhaftigkeitsuntersuchung der Bonusbank im Vergleich zum Fixlohnvertrag wird in diesem Abschnitt nicht vorgenommen. Die Erkenntnis aus dem Abschnitt 4.4.2, dass die Bonusbanknutzung durch das geringere Verlustrisiko steigt, hat sich als robust erwiesen. Daher sind neue Wirkungen durch eine Verlustentstehung in $t = 2$ nicht zu erwarten. Wird allerdings die periodische Vergütung als relevante Alternative zur Bonusbank herangezogen, so zeigen sich neue Vorteilhaftigkeitsunterschiede auf. Aus Kapitel 4.2.3 ist bekannt, dass es für den Vorteilhaftigkeitsvergleich der leistungsabhängigen Vergütungssysteme ausreicht, sich auf die

[236] Es wäre auch denkbar, die Vorteilhaftigkeit einer Bonusbank in Konstellationen der selektiven Anstrengung zu untersuchen. In diesen Fällen würde der Prinzipal abhängig von der Realisation des Cash-Flows $\tilde{x}_1$ entscheiden, ob der Agent sich in der zweiten Periode anstrengen soll oder nicht. Janocha (2014) kommt zu dem Ergebnis, dass die Vorteilhaftigkeit der Bonusbank bei selektiver Anstrengung des Agenten steigt. Meines Erachtens spielt dieser Fall in der Realität eine untergeordnete Rolle und wird daher vernachlässigt.

[237] Vgl. Niemann (2011), S. 16f.

[238] Die Liquiditätsbedingung $x_2^L + (1+i)\big((1-\tau_c)x_1^L + L\big) \geq 0$ wird hierzu als erfüllt angenommen. Sie stellt sicher, dass auch bei niedrigem Output in $t = 1$ ausreichend liquide Mittel bei Realisation von $x_2^L < 0$ vorliegen.

Zustände (x_1^H, x_2^H) und (x_1^H, x_2^L) zu beschränken, da bei Realisation des niedrigen Outputs in $t = 1$ die zustandsabhängigen Endvermögen ohnehin gleich hoch sind. Insbesondere wird die Endvermögensbestimmung im Zustand (x_1^H, x_2^L) durch die Verlustentstehung in $t = 2$ beeinflusst.

Im Anhang wird beispielhaft die Endvermögensermittlung bei der periodischen Vergütung im Zustand (x_1^H, x_2^L) mithilfe eines Finanzplans aufgezeigt, um die relevanten Fallkonstellationen zu identifizieren. In $t = 1$ ergibt sich keine Änderung im Vergleich zum Grundmodell. Der Gewerbeertrag ergibt sich als Differenz des Outputs und dem als Betriebsausgabe abziehbaren Entlohnungsteils. Überschüssiges Geldvermögen wird weiterhin am Kapitalmarkt angelegt. In $t = 2$ ist bei Realisation von $\tilde{x}_2 = x_2^L$ die Endvermögensermittlung davon abhängig, ob das zvE in $t = 2$ positiv oder negativ ist. Das zvE in t=2 ist bei einer periodischen Vergütung bzw. Bonusbankvereinbarung gegeben durch:

$$x_2^L + i\Big[(1-\tau_c)\big(x_1^H - w_1^H\big) - k\tau_c w_1^H + L\Big] \tag{54}$$

$$x_2^L + i\Big[(1-\tau_c)\big(x_1^H - (1-\alpha)w_1^{H,BB}\big) - k\tau_c(1-\alpha)w_1^{H,BB} + L\Big] \tag{55}$$

Interessant ist lediglich der Fall, in dem bei mindestens einer der leistungsabhängigen Entgelte ein negatives zvE entsteht. Ansonsten wären die Zinserträge aus der ersten Periode so hoch, dass keine steuerlichen Verluste in $t = 2$ entstehen. Dies entspricht jedoch dem bereits untersuchten Fall der vollständigen Steuererstattung im Grundmodell bzw. dem Gewinnfall II. Die möglichen Fallkonstellationen in $t = 2$ werden in der Tabelle 10 zusammengefasst.

Die Fallbezeichnungen entsprechen denjenigen bei Verlustentstehung in $t = 1$. Weil die Entlohnungszahlung der ersten Periode bei der Bonusbank mit $(1-\alpha)w_1^{H,BB} < w_1^H$ geringer ist, ist für bestimmte Parameterausprägungen ein Hybridfall II denkbar. In diesem liegt aufgrund der höheren Verzinsung der ersten Periode in $t = 2$ bei einer Bonusbank kein Verlust vor, bei einer periodischen Vergütung allerdings schon. Da der Hybridfall II nur für wenige Parameterkonstellationen gültig ist und der Gewinnfall II bereits ausgeschlossen wurde, wird in der Arbeit lediglich der Verlustfall II näher in den Fokus genommen. Steuerliche Verluste liegen im Verlustfall II in den Zuständen (x_1^j, x_2^L) vor.

Das Endvermögen des Prinzipals bei hohem Output in beiden Perioden ist unverändert durch

$$EV_{T=3}^{HH,BB} = \left(1+i_{\tau_c}\right)EV_{T=2}^{HH,BB} + \left(1+i_{\tau_c}\right)^2 L \tag{56}$$

gegeben. Verglichen mit dem mehrperiodigen Grundmodell muss das Endvermögen aus der Kooperation erneut unter Berücksichtigung der liquiden Mittel auf den Bewertungszeitpunkt $T = 3$ bezogen werden. Die Gleichung (56) trifft analog für die periodische Vergütung zu, wenn $EV^{HH,BB}$ durch $EV^{HH}\left(a^H\right)$ ersetzt wird. Wenn dagegen nach erfolgreicher erster Periode $(\tilde{x}_1 = x_1^H)$ in $t = 2$ der niedrige Output $x_2^L < 0$ auftritt, bewirken die steuerlichen Verlustverrechnungsbeschränkungen eine Endvermögensänderung. Für das zustandsabhängige Endvermögen gilt bei einer Bonusbankvereinbarung dann:

$$\begin{aligned} EV_{T=3}^{HL,BB} = {} & (1+i(1-\tau_c\theta))\Big[x_2^L + (1+i)\big((1-\tau_c)x_1^H \\ & -(1-\tau_c(1-k))(1-\alpha)w_1^{H,BB}\big) + L\Big] - \\ & \tau_c\theta\Big(x_2^L + i\Big[(1-\tau_c)x_1^H - (1-\tau_c(1-k)) \\ & (1-\alpha)w_1^{H,BB} + L\Big]\Big) \end{aligned} \tag{57}$$

Die ersten beiden Zeilen der Gleichung (57) sind die aus der zweiten Periode angelegten Zahlungsüberschüsse, die auf den Zeitpunkt $T = 3$ bewertet sind.[239] Die Zinserträge des Anlagebetrages unterliegen in $t = 3$ grundsätzlich dem Unternehmensteuersatz. Der fällige Steuersatz hängt jedoch von der zukünftigen Ertragslage θ ab. Die dritte Zeile der Gleichung repräsentiert den Steuererstattungsbetrag des aus $t = 2$ vorgetragenen Verlustes. Die Gleichung (57) ist analog für das zustandsabhängige Endvermögen $EV^{HL}\left(a^H\right)$ bei einer periodischen Vergütung erfüllt, wenn $(1-\alpha)w_1^{H,BB}$ durch w_1^H ersetzt wird.

Steuerinduzierte Wirkungen im Verlustfall II.

Es lässt sich zeigen, dass der Prinzipal die Bonusbank der periodischen Vergütung im Verlustfall II vorzieht, sofern[240]

$$\begin{aligned} & \delta\Big(\phi^H\left(1+i_{\tau_c}\right) + \left(1-\phi^H\right) \\ & \left(\frac{1+i(1-\tau_c\theta)}{1+i_{\tau_c}}(1+i) - \frac{i\tau_c\theta}{1+i_{\tau_c}}\right)\Big) - 1 \geq 0. \end{aligned} \tag{58}$$

Weiterhin ist die Nutzung der Bonusbank durch das Abwägen des positiven Zinseffektes und des negativen Einflusses der kurzfristigen Konsumpräferenz beim Manager gegeben. Analog zum Grundmodell ist für $i = 0$ und $\delta = 1$ die Differenz gleich Null, während der Prinzipal für $i = 0$ und $\delta < 1$ die periodische Vergütung vorzieht.

Können Erträge am Kapitalmarkt erwirtschaftet werden $(i > 0)$ und ist der Unternehmensteuersatz strikt positiv $(\tau_c > 0)$, so begünstigt die mögliche Verlustentstehung mit der Wahrscheinlichkeit $1-\phi^H > 0$ die Nutzung der Bonusbank, zumal für den Klammerausdruck gilt:[241]

$$\frac{1+i(1-\tau_c\theta)}{1+i_{\tau_c}}(1+i) - \frac{i\tau_c\theta}{1+i_{\tau_c}} > \left(1+i_{\tau_c}\right). \tag{59}$$

Die asymmetrische Verlustbehandlung ist ursächlich für diesen positiven Zinseffekt, weil sich die Vorteilhaftigkeitsbedingung bei nahezu ausgeschlossenem Cash-Flow x_2^L durch $\phi^H \to 1$ an die Bedingung im Grundmodell annähert.[242]

[239] Zahlungsüberschüsse liegen wegen der erfüllten Liquiditätsbedingung immer vor.

[240] Siehe Anhang.

[241] Der Beweis erfolgt im Anhang.

[242] Wenn die Bedingung (59) mit Gleichheit erfüllt wäre, so entspreche sie ebenfalls der Vorteilhaftigkeitsbedingung (37) aus dem Grundmodell.

Fall-bezeichnung	(x_1^H, x_2^L) und PV	(x_1^H, x_2^L) und BB	(x_1^L, x_2^L) und BB/PV	Bedingung
Gewinnfall II	*Gewinn*	*Gewinn*	*Gewinn*	$x_2^L + i\big((1-\tau_c)x_1^L + L\big) \geq 0$
Hybridfall II	*Verlust*	*Gewinn*	*Verlust*	(54) < 0 und (55) ≥ 0
Verlustfall II	Verlust	Verlust	Verlust	$x_2^L + i\big[(1-\tau_c)x_1^H - (1-\tau_c(1-k))(1-\alpha)w_1^{H,BB} + L\big] < 0$

Tabelle 10: Fälle bei möglicher Verlustentstehung in $t = 2$ bezogen auf das zvE in $t = 2$ und Implementierung der permanenten Anstrengung (Eigene Darstellung)

Die geringere Endvermögensbenachteiligung der Bonusbank entsteht durch den größeren Kapitalmarktanlagebetrag aufgrund des einbehaltenen Bonusbankanteils $\alpha w_1^{H,BB}$ zum Ende von $t = 1$. Dieser führt in $t = 2$ zu einem geringeren Verlustvortragsbetrag, der mit einem geringeren Zinsverlust einhergeht. Da die linke Seite von (59) in θ sinkt, steigt die relative Vorteilhaftigkeit der Bonusbank zusätzlich durch weitere Verlustausgleichsrestriktionen.

Als Instrument zur Entscheidungsfindung hinsichtlich der Wahl einer der leistungsabhängigen Vergütungen wurde bereits der kritische Bruttozinssatz i_{krit} identifiziert. Änderungen der kritischen Zinssätze im Vergleich zum Grundmodell geben Aufschluss über die Signifikanz des Bonusbankvorteils aus der differenzierten Behandlung von Gewinnen und Verlusten. Die kritischen Zinssätze i_{krit} lassen sich nicht mehr explizit ermitteln, sondern sind implizit durch die Bedingung (58) definiert. Die absoluten Zinsdifferenzen werden in Abhängigkeit der relevanten Parameter θ, τ_c und δ in der Tabelle 11 dargestellt.[243] Für alle variierten Werte in der Tabelle stellt die nachfolgende Datenkonstellation den Verlustfall II sicher.

Da die Bonusbank aufgrund des Zinsvorteils aus der Verlustentstehung in $t = 2$ begünstigt wird, sind die Änderungen der kritischen Zinssätze in der Tabelle 11 negativ. Bei vollständigem Verlustausgleich ($\theta = 1$) liegen die Minderungen der kritischen Zinssätze zwischen 0% und 0,23% und sind als gering einzustufen, auch wenn die Entscheidungsrelevanz der Änderungen ex ante nicht ausgeschlossen werden kann. Für $\delta = 0,95$ und $\tau_c = 29,83\%$ sinkt der Schwellenwert i_{krit} bspw. von 7,50% auf 7,45%. Werden in $t = 3$ weitere Verlustverrechnungsbeschränkungen ($\theta = 0,5$) berücksichtigt, so liegen die Änderungen zwischen -0,15% und -2,32% bzw. für $\delta = 0,95$ und $\tau_c = 29,83\%$ gerade bei einem Prozentpunkt. Damit fallen die Änderungen beträchtlicher aus als bei einem vollständigen Verlustausgleich.

Die angegebenen kritischen Zinssätze müssen hinsichtlich ihrer Anwendbarkeit mit Vorsicht interpretiert werden. Der Fall $\theta = 0,5$ sensibilisiert den Blick für die Tatsache, dass das zukünftige Verlustabzugspotential die Schwellenwertzinssätze stark beeinflusst. Gleichzeitig ist aber unklar, wie genau die Ertragsprognose eines Unternehmens in konkrete Werte von θ zu transferieren ist. Der Fall $\theta = 1$ symbolisiert ein Steuersystem, in dem unbeschränkte Verlustvorträge vorgesehen sind und genügen Ausgleichspotenzial vorhanden ist. Die prozentualen Änderungen für $\theta = 1$ sind in zweifacher Hinsicht als Minimaländerungen anzusehen. Zum einen verstärken sich Schwellenwertsenkungen mit zunehmenden Verlustverrechnungsbeschränkungen wie bspw. der Mindestbesteuerungsquote im deutschen Steuerrecht.[244] Zum anderen werden Bonusbankvereinbarungen typischerweise über längere Zeiträume als die modellierten zwei Perioden geschlossen. Bei Einbehaltung von Anteilen α in jeder Periode des längeren Zeitraums stellen sich erheblichere Zinseffekte durch die Einbehaltung ein, die in Kombination mit Verlustentstehungsmöglichkeiten in sämtlichen Perioden eine größere Wirkung entfalten. Daher ist die geringe Auswirkung auf den kritischen Zinssatz auch der Zwei-Jahres-Kooperation geschuldet.

Im Verlustfall II wird insgesamt deutlich, dass das Endvermögen des Prinzipals bei der Sofortauszahlung von Boni gegenüber einer Bonusbankvereinbarung stärker durch potentielle Verluste geschmälert wird. Dies liegt jedoch nicht an dem in der Praxis diskutierten Vorteil der Bonusbank, eine direkte Verlustbeteiligung des Managers zu bewirken. In dem Prinzipal-Agenten-Modell zeigt sich, dass durch die Haftungsbeschränkung beim Manager eine direkte Verlustbeteiligung ausgeschlossen ist. Der einbehaltene Anteil α bewirkt dies nur scheinbar, weil zur Anreizsetzung die Entlohnungszahlung $w_1^{H,BB}$ höher angesetzt werden muss.

Der Vorteil der Bonusbank bei potentieller Verlustmöglichkeit in $t = 2$ liegt in der langfristig glättenden Bonusausschüttung. Durch sie schlägt die asymmetrische Verlustbehandlung nicht so extrem durch wie bei einer periodischen Vergütung. Die hiesige Zwei-Perioden-Struktur führt aber dazu, dass die effektive Endvermögensbenachteiligung marginal ist, weil die Wirkung der differierenden Entlohnungszahlungen sich lediglich über die Verzinsung aus $t = 1$ auf die Folgeperioden auswirkt.[245] Unter Beachtung von typischer-

[243] Die numerische Ermittlung erfolgte mit dem Excel-Solver. Die Angaben sind auf zwei Nachkommastellen gerundet.

[244] Von dem Sockelbetrag des § 10d EStG und Verlustrückträgen sei in der Interpretation abgesehen, da diese in großen Kapitalgesellschaften von geringer Bedeutung sind.

[245] Für die angegebene Datenkonstellation sowie $\theta = 0.5, i = 0.1, \tau_c = 0,3$ und $\delta = 0,95$ beträgt der effektive Endvermögensvorteil nur 0,08 GE.

x_t^H	x_1^L	x_2^L	L	ϕ_t^H	ϕ_t^L	v^H	α	$\tau_{inc,t}$	τ_d	k
150	0	-100	100	0,65	0,45	6	0,5	0,42	0,25	0

Hebesatz	300%	350%	400%	450%	500%
τ_c	26,33%	28,08%	29,83%	31,58%	33,33%
	Änderungen von i_{krit} für $\theta = 1$				
$\delta = 0{,}91$	-0,15%	-0,17%	-0,18%	-0,20%	-0,23%
$\delta = 0{,}93$	-0,09%	-0,10%	-0,11%	-0,12%	-0,14%
$\delta = 0{,}95$	-0,04%	-0,05%	-0,05%	-0,06%	-0,07%
$\delta = 0{,}97$	-0,00%	-0,00%	-0,02%	-0,02%	-0,02%
$\delta = 0{,}99$	-0,00%	-0,00%	-0,00%	-0,00%	-0,00%
	Änderungen von i_{krit} für $\theta = 0,5$				
$\delta = 0{,}91$	-1,56%	-1,73%	-1,91%	-2,11%	-2,32%
$\delta = 0{,}93$	-1,18%	-1,31%	-1,44%	-1,59%	-1,74%
$\delta = 0{,}95$	-0,82%	-0,90%	-1,00%	-1,10%	-1,21%
$\delta = 0{,}97$	-0,47%	-0,53%	-0,58%	-0,64%	-0,70%
$\delta = 0{,}99$	-0,15%	-0,17%	-0,19%	-0,21%	-0,23%

Tabelle 11: Änderungen der kritischen Bruttozinssätze in Abhängigkeit von der Konsumpräferenzrate, dem Gewerbesteuerhebesatz und der Ertragslage θ im Verlustfall II (Eigene Darstellung)

weise länger andauernden Bonusbankvereinbarungen sowie Verlustverrechnungsbeschränkungen $\theta < 1$ in *jeder* Periode sind größere Endvermögensbenachteiligungen der periodischen Vergütung zu erwarten. Dies gilt insbesondere, wenn die Annahme des binären Cash-Flows aufgegeben wird und der Prinzipal genauere Entlohnungsabstufungen als w^L bzw. w^H festlegen kann.

In dem Modellrahmen der Arbeit bleibt ein größerer Endvermögensvorteil der Bonusbank dadurch unsichtbar, dass bei Verlustentstehung $x_t^L < 0$ keine Entlohnungszahlungen ($w_t^L = w_t^{L,BB} = 0$) vorliegen und die Verlustsituation durch diese Vergütungen nicht weiter forciert wird. Die Einbehaltung von Anteilen α entschärft damit grundsätzlich die steuerliche Verlustentstehung. Zwar geht diese Glättung mit einer höheren, erwarteten Entlohnungszahlung einher ($w_1^{H,BB} > w_1^H$), falls der Manager die gewünschten Ziele erreicht. Dieser Nachteil kann jedoch unter Umständen durch Zinseffekte wieder ausgeglichen werden.

4.5. Modellkritik

Die nachgewiesenen, steuerlichen Effekte sind stark abhängig von den zugrunde liegenden Modellannahmen. Bevor die Ergebnisse tatsächlich als zusätzliche Informationen für die Gestaltung eines effizienten Anreizsystems gewertet werden können, sind aus der Prämissensetzung einhergehende Limitationen zu prüfen.

Die vorliegende Arbeit geht davon aus, dass der Output lediglich zwei Ausprägungen annehmen kann.[246] In der Praxis werden sicherlich nicht nur zwei mögliche Szenarien für die zukünftige Unternehmensentwicklung in Betracht gezogen. Die Entlohnung des Managers wird ebenso nicht nur an zwei Realisationen festgemacht und es gibt mehrere Abstufungen in den variablen Vergütungen. Insofern mindert der binäre Output den Erklärungsgehalt des Modells. Das Problem bei einer Vielzahl möglicher Cash-Flows ist die größere Anzahl möglicher Zustände, für die die Endvermögen, ggf. unter Beachtung von Verlustvorträgen, separat ermittelt werden müssen. Insofern ist fraglich, ob die exakte Erfassung von Zeiteffekten in einem erweiterten Rahmen möglich ist.[247]

Eine streng mehrperiodige Modellerweiterung erfordert die explizite Einbeziehung des Kapitalmarktes auf Managerebene, die im vorliegenden Modell nicht vorgenommen wird.[248] Kapitalmarktaktivitäten des Agenten werden hier allenfalls implizit über eine variierende Konsumpräferenzrate abgebildet. Die Schwierigkeit besteht darin, dass die Leistungserbringung des Agenten von den Kapitalmarktaktivitäten abhängig ist und dadurch eine Periodenabhängigkeit in der Anstrengungswahl des Agenten geschaffen wird. Zur analytischen Lösbarkeit und Komplexitätsreduktion erweist

[246] Das binäre Anstrengungsniveau ist unkritisch, da abweichende Ergebnisse hieraus nicht zu erwarten sind. Siehe dazu auch die Ausführungen im Anhang.

[247] Vgl. ähnlich Niemann (2011), S. 19.

[248] Eine derartige Erweiterung steht in der Literatur zu steuerlichen Anreizwirkungen noch aus.

sich die Annahme als notwendig, wenngleich abweichende Steuerwirkungen nicht auszuschließen sind.

Die Modellierung der Bonusbank ist in Anbetracht der tatsächlichen Ausgestaltung in der Praxis stark vereinfachend. Bonusbankvereinbarungen werden über deutlich längere Zeiträume geschlossen, sodass die zweiperiodige Betrachtung restriktiv ist. Dadurch können Zinswirkungen nicht genau abgebildet und effektive Endvermögensvorteile von Bonusbanken lediglich erahnt werden. Dies gilt speziell für die Analyse unter Einbeziehung steuerlicher Verluste auf Unternehmensebene. Darüber hinaus schließt die Haftungsbeschränkung des Agenten den Verdienst negativer Beträge aus, die mit positiven Beträgen in zukünftigen Perioden verrechnet werden. Wie aus der Funktionsbeschreibung der Bonusbank in Kapitel 2.2.3 deutlich wird, macht dies aber gerade die Bonusbank aus. Alles in allem muss die vereinfachte Abbildung der Bonusbank als Kompromiss zur expliziten Untersuchung der Anreizwirkungen im Prinzipal-Agenten-Modell gesehen werden.

In Zusammenhang mit der Modellierung der steuerlichen Verlustvorschriften sind ebenfalls Einschränkungen zu beachten. Besonders die Mindestbesteuerung als zusätzliche Verlustverrechnungsbeschränkung wird nicht konkret modelliert, sondern von der Wirkrichtung über den Parameter θ erfasst. Dies führt dazu, dass resultierende Zinsnachteile in abgeschwächter Form zur Geltung kommen. Darüber hinaus ist die dritte Periode als künstliches Verlustausgleichskonstrukt zu sehen, da ein deterministischer Cash-Flow $x_3 > 0$ in einem Modell unter Unsicherheit nicht zu rechtfertigen ist. Einzig die fehlende Alternative in der Literatur und die Notwendigkeit der Bewertung von verbleibenden Verlustvorträgen legitimiert dieses Vorgehen.

Bei der Analyse von Verlustentstehungsmöglichkeiten wurde stets von vorhandenen liquiden Mitteln L am Ende von $t = 1$ ausgegangen. Im Rahmen einer Grenzbetrachtung können diese liquiden Mittel in Kapitalgesellschaften nur aus anderen Unternehmensbereichen stammen.[249] Weil Bonusbanken eher in höheren Hierarchieebenen genutzt werden, beeinflusst der Manager durch seine Handlung die Rückflüsse der Gesamtunternehmung. Daher ist die Annahme, dass die liquiden Mittel aus einem anderen Unternehmensbereich kommen, für den Fall der Realisation eines Cash-Flows $x_t^L < 0$ nicht realistisch. In diesem Fall muss streng genommen eine Fremdfinanzierung von $x_t^L < 0$ erfolgen. Dann sind die Bemessungsgrundlagen für die Körperschaftsteuer und die Gewerbesteuer nicht identisch, zumal bei der Ermittlung des Gewerbeertrages ein Viertel der fälligen Darlehenszinsen, die bei der Ermittlung des zvE abgezogen werden, wieder hinzuzurechnen ist. Dies stellt eine zusätzliche Endvermögensreduktion beim Prinzipal dar. Diese Regelung benachteiligt tendenziell den Fixlohnvertrag aufgrund der wahrscheinlicheren Verlustentstehung, da die Managervergütungen zum Ende der Perioden anfallen und positive Vergütungen ausschließlich bei hohem Output gezahlt werden.

[249] Vgl. ähnlich König und Wosnitza (2004), S. 91.

Andere Steuerwirkungen können resultieren, falls die Entlohnungszahlung bereits zum Anfang einer Periode erfolgt. Bei Verlustentstehungsmöglichkeit in $t = 1$ und Realisation des Zustandes (x_1^L, x_2^H) ist in dem Fall eine Fremdkapitalaufnahme in Höhe von $-x_1^L + w_2^H$ nötig, während beim Fixlohnvertrag eine Kapitalaufnahme nie über den negativen Cash-Flow hinausgeht. Dies stellt einen Endvermögensnachteil in diesem Zustand bei den leistungsabhängigen Vergütungen dar. Für eine genaue Untersuchung der effektiven Endvermögensnachteile müsste diese differenzierte Behandlung der Fremdfinanzierung bei der Körperschaftsteuer und Gewerbesteuer explizit in den Modellrahmen integriert werden. Vermutlich überwiegt der Nachteil auf Seiten des Fixlohnvertrages ähnlich wie im Hybridfall I.

Insgesamt ist festzustellen, dass durch die Prämissensetzung eine monetäre Quantifizierung der tatsächlichen Steuerwirkungen nur schwer möglich ist. Gleichwohl erfolgt die Identifikation von Wirkrichtungen, was letztlich das zentrale Anliegen dieser Arbeit ist.

5. Fazit

Vor dem Hintergrund der regelmäßig in der Kritik stehenden Managementvergütungen und der gesetzten Fehlanreize vor der Weltwirtschaftskrise durch Bonuszahlungen, deren Höhe an kurzfristige Unternehmenskennzahlen gekoppelt wurde, wird in dieser Arbeit die Bonusbank als Anreizinstrument analysiert. Die Untersuchung erfolgt unter simultaner Berücksichtigung des modellendogenen Anreizproblems zwischen einem Manager und einem Anteilseigner sowie von steuerlichen Einflüssen, während thematisch ähnliche Arbeiten sich lediglich auf einen der genannten Aspekte beschränken. Zudem wird erstmals der Einfluss von Ertragsteuern auf die Anreizsetzung in einem mehrperiodigen Prinzipal-Agenten-Modell in den Blick genommen.

Zur formalen Ergebnisherleitung wird ein Modell mit binärem Anstrengungs- und Outputniveau sowie einem risikoneutralen Agenten, der haftungsbeschränkt ist, genutzt. Neben der Möglichkeit, die Bonusbankvereinbarung zu verwenden, besitzt der Prinzipal mit dem periodisch ausgezahlten Bonus und dem Fixlohnvertrag zwei weitere Alternativen. Jede dieser Vergütungsoptionen kann dem Prinzipal das maximale Endvermögen liefern.

Reale Steuersysteme beeinflussen die Vergütungswahl und damit den Anreizkonflikt auf unterschiedlichen Ebenen. Zur Abbildung der allumfassenden Steuerwirkungen hinsichtlich der Entlohnungsalternativen werden daher die zentralen Steuervorschriften integriert, die im Rahmen der Grenzbetrachtung relevant sind und in der Literatur bereits identifiziert wurden. Dazu zählen insbesondere die Unternehmensteuer auf Kapitalgesellschaftsebene und die Einkommensteuer beim Manager. Auf Unternehmensebene wird zudem der Einfluss der Nichtabzugsfähigkeit der Managerentlohnung als Betriebsausgabe sowie die asymmetrische Behandlung von steuerlichen Gewinnen und Verlusten in Bezug auf die Vorteilhaftigkeit der Vergütungssysteme untersucht.

Unter Ausschluss steuerlicher Verluste auf Ebene der Kapitalgesellschaft ist festzustellen, dass steuerliche Vorschriften die Nutzung der Bonusbank gegenüber einem Fixlohnvertrag benachteiligen. Die Entlohnungszahlung bei einer Bonusbank zur Setzung von Leistungsanreizen übersteigt diejenige bei einem fixen Gehalt, was die vom Prinzipal zu tragende Einkommensteuerbelastung des Managers sowie die Kosten einer (potentiellen) Nichtabzugsfähigkeit als Betriebsausgabe erhöht. Diese Steuerwirkungen gelten im übrigen unabhängig davon, ob die Einkommensteuer des Managers periodenabhängig ist und eine mehrperiodige Betrachtung vorgenommen wird oder nicht. Werden steuerliche Verluste auf Unternehmensebene berücksichtigt, so zeigt sich ein umgekehrtes Bild. Durch die erhöhte Wahrscheinlichkeit der Verlustentstehung bei der leistungsunabhängigen Entlohnung ist der Zinsverlust aus den Verlustverrechnungsbeschränkungen größer, sodass die effektive Endvermögensbenachteiligung aus Prinzipalsicht durch die Bonusbank geringer ist. Dieses Resultat hält sämtlichen Parametervariationen stand und ist deshalb interessant, weil erhöhte Entlohnungszahlungen die Verlustsituation zwar verstärken, die Gesamtbeurteilung aber nicht verändern.

Wird bei der Untersuchung der Bonusbank die periodische Vergütung als relevante Alternative festgelegt, so zeigt sich ein heterogenes Bild. Der Unterschied dieser beiden leistungsabhängigen Vergütungen ist gerade der in die Bonusbank eingestellte Vergütungsanteil, der in Abhängigkeit von dem Unternehmensergebnis in der zweiten Periode ausgezahlt wird. Im Grundmodell ergibt sich durch die Einbehaltung der Vorteil, dass auf Unternehmensebene ein zusätzlicher Zinsertrag generiert werden kann. Der effektive Zinsertrag sinkt mit zunehmendem Unternehmensteuersatz, zumal Zinserträge steuerbare Einnahmen darstellen. Durch die Einbehaltung des Vergütungsanteils ergibt sich allerdings der Nachteil, dass der Agent den sofortigen Erhalt der Entlohnungszahlung vorzieht. In Abhängigkeit vom Unternehmensteuersatz und der Konsumpräferenz des Agenten lassen sich kritische Zinssätze als Entscheidungskriterium für den Prinzipal ermitteln.

Sofern die Grenzeinkommensteuersätze des Managers über die Perioden nicht konstant sind, werden diese im Gegensatz zum mehrperiodigen Grundmodell beim Vergleich mit der periodischen Vergütung entscheidungsrelevant. Während höhere Steuersätze den Gebrauch von leistungsabhängigen Vergütungen generell hemmen, verringern höhere Grenzsteuersätze in früheren (bzw. späteren) Perioden das Endvermögen beim Einsatz der periodischen Vergütung (bzw. der Bonusbank). Da die tarifliche Einkommensteuer im deutschen Steuerrecht durch im zu versteuernden Einkommen steigende Grenzsteuersätze geprägt ist, können einbehaltene Bonusanteile einen Progressionsvorteil bewirken, während bei erfolgreicher Unternehmensentwicklung durch die geballte Ausschüttung Progressionsnachteile entstehen können.

Der Einfluss der asymmetrischen Behandlung von steuerlichen Gewinnen und Verlusten wird im Vergleich zur periodischen Vergütung lediglich bei der Verlustentstehung in der zweiten Periode sichtbar. Weil bei realisierten Verlusten eine Bonusbank durch die geringere Entlohnungszahlung der Vorperiode die Verlustsituation abschwächt, reduzieren Verlustverrechnungsbeschränkungen das Endvermögen des Prinzipals in geringerem Maße. Sicherlich muss konstatiert werden, dass der steuerliche Effekt in diesem Modellrahmen nur selten einen Vorteilhaftigkeitswechsel induziert, da der erwartete Endvermögenseffekt relativ gering ist. Jedoch ist dies auch dem vereinfachenden Zwei-Perioden-Modell geschuldet, der den Glättungseffekt der Bonusbank und effektive Zinswirkungen nicht in Gänze herausheben kann.

In Anbetracht der Tatsache, dass Bonusbanken in der Praxis über längere Zeiträume als den Zwei-Jahres Zeitraum vereinbart werden, ist durch die Bonusbank ein aus steuerlicher Sicht vorteilhafter Glättungseffekt in zweierlei Hinsicht zu erwarten. Zum einen kann die Bonusbank die Progressionswirkung durch die Einstellung von Guthabensbeträgen in die Bonusbank abmildern. Zum anderen entschärft die Bonusbank die Verlustentstehung auf Unternehmensebene, falls schlechte Unternehmensentwicklungen im Bereich des Möglichen sind. Sicherlich sind diese Interpretationen als Tendenzaussagen zu werten, jedoch liefert die Analyse dafür einige Anhaltspunkte. Eine quantitative Bewertung dieser Effekte ist in dem Prinzipal-Agenten-Modell schwierig, wenngleich kritische Zinssätze auch bei periodenabhängiger Einkommensteuer und bei Verlustentstehung in der zweiten Periode ermittelt werden.

Die Ergebnisse können zum Anlass genommen werden, steuerpolitische Maßnahmen zu erörtern. Sofern der Gesetzgeber bspw. die Nutzung von Bonusbanken nicht nur empfehlen, sondern konkret fördern möchte, bietet sich ihm durch eine Anpassung des § 34 EStG ein mögliches Instrument. Das Zustandekommen eines Progressionsnachteils durch mehrjährig akkumulierte Auszahlungen aus einer Bonusbank könnte verhindert werden, wenn eine Qualifikation des entsprechenden Endguthabenanteils als Vergütung aus mehrjähriger Tätigkeit möglich wäre. Paradoxerweise scheitert die Anwendung der begünstigten Fünftelmethode nach jetziger Rechtslage ausschließlich an der leistungsabhängigen Komponente der Bonusbank, die der Gesetzgeber eigentlich fördern will. Ein positiver Nebeneffekt dieser Maßnahme wäre zudem, dass die Implementierung der hohen, im Regelfall effizienten Leistung des Agenten im Vergleich zum Fixlohnvertrag gefördert würde. In diesem Zusammenhang wird in dieser Arbeit ebenfalls kein Indiz dafür gefunden, dass eine beschränkte Abzugsfähigkeit des Managergehalts aus anreiztheoretischen Gesichtspunkten eine erwägenswerte Maßnahme ist.

Hinsichtlich der Vorteilhaftigkeit von Bonusbanken bleiben weiterhin Fragen offen. So könnte in Folgearbeiten untersucht werden, welche Anreize Bonusbankvereinbarungen bei einem risikoaversen Agenten schaffen. Dabei ist generell denkbar, dass die glättenden Einkommensströme aus der Bonusbank die Nutzung einer solchen begünstigen, während die doppelt leistungsabhängige Komponente aufgrund der riskanteren Einkommenserzielung mit der Zahlung einer Risikoprämie verbunden ist. Außerdem sind aus Verlust-

verrechnungsbeschränkungen entstehende Zeiteffekte bei risikoaversem Agenten noch nicht untersucht worden. Das Instrumentarium könnte mit dem zweiperiodigen Modell dieser Arbeit gegeben sein. Es bleibt abzuwarten, ob eine sinnvolle Lösbarkeit trotz der ohnehin komplexen Lösungsermittlung bei risikoneutralen Akteuren realisierbar ist.

Literatur

Aders, C. & Hebertinger, M. (2003). Shareholder-Value-Konzepte: Eine Untersuchung der DAX100-Unternehmen. In W. Ballwieser, P. Wesner & KPMG (Hrsg.), *Value based management*. Frankfurt am Main.

Arnold, M. C. & Gillenkirch, R. M. (2007). Leistungsanreize durch Aktien oder Optionen? Eine Diskussion des State of the Art. *Journal of Business Economics*, *77* (1), 75–99.

Bauer, T., Kourouxous, T. & Krenn, P. (2018). Taxation and agency conflicts between firm owners and managers: a review. *Business Research*, *11* (1), 33–76.

Bebchuk, L. A. & Fried, J. M. (2003). Executive compensation as an agency problem. *Journal of Economic Perspectives*, *17* (3), 71–92.

Becker, F. G. (1990). *Anreizsysteme für führungskräfte: Möglichkeiten zur strategisch-orientierten steuerung des managements*. Stuttgart: Poeschel.

Becker, F. G. (2007). *Organisation der unternehmungsleitung: Stellgrößen der leitungsorganisation*. Stuttgart: Kohlhammer.

Becker, F. G. (2015). *Grundlagen der unternehmungsführung: Einführung in die managementlehre* (3., neu bearb. Aufl. Aufl.). Berlin: Schmidt.

Becker, F. G. & Ostrowski, Y. (2012). Materielle Anreizsysteme für Führungskräfte. *WiSt-Wirtschaftswissenschaftliches Studium*, *41* (10), 526–531.

Benmelech, E., Kandel, E. & Veronesi, P. (2010). Stock-based compensation and CEO (dis) incentives. *The Quarterly Journal of Economics*, *125* (4), 1769–1820.

Berle, A. A. & Means, G. C. (1932). *The modern corporation and private property*. New York: MacMillan.

Bhagat, S. & Bolton, B. (2014). Financial crisis and bank executive incentive compensation. *Journal of Corporate Finance*, *25*, 313–341.

Bierbaum, J. (2002). Repeated Moral Hazard under Limited Liability. *SSRN Electronic Journal*.

Blohm, H., Lüder, K. & Schäfer, C. (2012). *Investition: Schwachstellenanalyse des Investitionsbereichs und Investitionsrechnung* (10., bearb. und aktualisierte Aufl. Aufl.). München: Vahlen. Zugriff auf http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=678744 doi: 10.15358/9783800639380

CDU/CSU & SPD. (2018). *Koalitionsvertrag zwischen CDU, CSU und SPD: 19. Legislaturperiode.* Berlin. Zugriff auf https://www.bundesregierung.de/resource/blob/975226/847984/5b8bc23590d4cb2892b31c987ad672b7/2018-03-14-koalitionsvertrag-data.pdf?download=1

Core, J. E., Guay, W. R. & Larcker, D. F. (2003). Executive equity compensation and incentives: A survey. *Federal Reserve Bank of New York Economic Policy Review*, *9* (1), 27–50.

Crasselt, N. & Fründ, H. P. (2014). Managementvergütung mit Bonusbanken. *WiSt-Wirtschaftswissenschaftliches Studium*, *43* (3), 165–170.

Crezelius, G. (2018). § 5 estg. In P. Kirchhof (Hrsg.), *Einkommensteuergesetz* (S. 425–489). Köln: Verlag Dr. Otto Schmidt.

Dierkes, S. & Schäfer, U. (2008). Das mehrperiodige LEN-Modell mit additiv separabler Nutzenfunktion. *WiSt-Wirtschaftswissenschaftliches Studium*, *37* (5), 238–244.

Döscher, T. (2014). *Stand und Entwicklung der Vorstandsvergütung in Deutschland und Europa: Eine Analyse der Ausgestaltung der Vergütungssysteme in Deutschland, Österreich, Frankreich und Großbritannien und deren Implikationen für die Forschung zur Vorstandsvergütung*.

Ewert, R. & Niemann, R. (2012). Steuern in Agency-Modellen: Mehrperioden-und Multi-Task-Strukturen. *Arqus-Diskussionsbeiträge zur quantitativen Steuerlehre*, 1–46.

Fama, E. F. & Jensen, M. C. (1983). Agency problems and residual claims. *The Journal of Law and Economics*, *26* (2), 327–349.

Fellingham, J. C. & Wolfson, M. A. (1985). Taxes and risk sharing. *Accounting Review*, 10–17.

Frenz, P., Schulz, S. & Petry, T. (2011). Vergütung - Dem Esel wird die Möhre höher gehängt: Inwieweit hat die Finanz- und Wirtschaftskrise eine stärkere Veränderungsdynamik bei variablen Vergütungssystemen bewirkt? *Personalwirtschaft*, *38* (4), 58–60.

Friedl, G., Hölz, C., Beck, D., Heilmann, C. & Pflugmann, F. (2018). Studie zur Vergütung der Vorstände in den DAX-und MDAX-Unternehmen im Geschäftsjahr 2017. , 1–10.

Frydman, C. & Jenter, D. (2010). CEO compensation. *Annual Review of Financial Economics*, *2* (1), 75–102.

Gillenkirch, R. (1997). *Gestaltung optimaler Anreizverträge: Motivation, Risikoverhalten und beschränkte Haftung* (Bd. 82). Wiesbaden: Gabler Verlag.

Grossman, S. J. & Hart, O. D. (1983). An Analysis of the Principal-Agent Problem. *Econometrica: Journal of the Econometric Society*, *51* (1), 7–45.

Halperin, R. M., Kwon, Y. K. & Rhoades-Catanach, S. C. (2001). The impact of deductibility limits on compensation contracts: A theoretical examination. *Journal of the American Taxation Association*, *23* (s-1), 52–65.

Holmström, B. (1979). Moral hazard and observability. *Bell Journal of Economics*, *10* (1), 74–91.

Hostettler, S. & Stern, H. J. (2004). *Das Value Cockpit: Sieben Schritte zur wertorientierten Führung für Entscheidungsträger* (1. Aufl. Aufl.). Weinheim: Wiley-VCH.

Hundsdoerfer, J., Kiesewetter, D. & Sureth, C. (2008). Forschungsergebnisse in der Betriebswirtschaftlichen Steuerlehre—eine Bestandsaufnahme. *Journal of Business Economics*, *78* (1), 61–139.

Janocha, M. (2014). *Vergütung mittels Bonusbanken: eine agency-theoretische Perspektive*. Springer-Verlag.

Jasper, T. (1995). *Entlohnungssysteme und Einkommensbesteuerung: Risiko, Anreize und Steuerarbitrage* (Bd. 16). Hamburg: S und W, Steuer- und Wirtschaftsverlag.

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.

Jensen, M. C. & Murphy, K. J. (1990). Performance pay and top-management incentives. *Journal of Political Economy*, *98* (2), 225–264.

Jost, P.-J. (2001a). Die Prinzipal- Agenten- Theorie im Unternehmenskontext. In P.-J. Jost (Hrsg.), *Die prinzipal-agenten-theorie in der betriebswirtschaftslehre* (S. 11–43). Stuttgart: Schäffer-Poeschel.

Jost, P.-J. (2001b). Vorwort. In P.-J. Jost (Hrsg.), *Die prinzipal-agenten-theorie in der betriebswirtschaftslehre* (S. 1–6). Stuttgart: Schäffer-Poeschel.

Kiesewetter, D. & Dietrich, M. (2007). Ein Standardmodell für Investitionsentscheidungen in Kapitalgesellschaften. *WiSt-Wirtschaftswissenschaftliches Studium*, *36* (5), 235–244.

König, R. & Wosnitza, M. (2004). *Betriebswirtschaftliche Steuerplanungs- und Steuerwirkungslehre*. Heidelberg: Physica-Verlag Heidelberg.

Kräkel, M. (2007). *Organisation und Management* (3. Aufl.). Tübingen: Mohr Siebeck.

Krenn, P. (2017). The impact of taxes on competition for CEOs. *European Accounting Review*, *26* (3), 503–530.

Laffont, J.-J. & Martimort, D. (2009). *The theory of incentives: the principal-agent model*. Princeton University Press.

Lambert, R. A. (1983). Long-term contracts and moral hazard. *The Bell Journal of Economics*, 441–452.

Laux, H. (2003). *Wertorientierte Unternehmensführung und Kapitalmarkt: Fundierung von Unternehmenszielen und Anreize für ihre Umsetzung*. Berlin, Heidelberg and s.l.: Springer Berlin Heidelberg.

Martens, K.-P. (2005). Die Vorstandsvergütung auf dem Prüfstand. *Zeitschrift für das gesamte Handelsrecht und Wirtschaftsrecht*, *169*, 124–154.

Mehrmann, A. & Sureth-Sloane, C. (2017). Tax loss offset restrictions and biased perception of risky investments. *Arqus-Diskussionsbeiträge zur quantitativen Steuerlehre*, 1–57.

Meißner, F., Schneider, G. & Scholze, A. (2014). Vergleichende Analyse verschiedener Ausprägungen einer erhöhten Besteuerung von Bonuszahlungen. *Schmalenbachs Zeitschrift für betriebswirtschaftliche Forschung*, *66* (3), 212–241.

Mellinghoff, R. (2018). § 34 EStG. In P. Kirchhof (Hrsg.), *Einkommensteuergesetz* (S. 1861–1885). Köln: Verlag Dr. Otto Schmidt.

Milgrom, P. R. (1981). Good news and bad news: Representation theorems and applications. *The Bell Journal of Economics*, 380–391.

Miller, M. H. & Scholes, M. S. (1982). Executive compensation, taxes, and incentives. In W. F. Sharpe & C. M. Cootner (Hrsg.), *Financial economics* (S. 179–201). Englewood Cliffs, NJ: Prentice-Hall.

Niemann, R. (2004). Investitionswirkungen steuerlicher Verlustvorträge-Wie schädlich ist die Mindestbesteuerung? *Journal of Business Economics*, *74* (4), 359–384.

Niemann, R. (2007). Risikoübernahme, Arbeitsanreiz und differenzierende Besteuerung. *Arqus-Diskussionsbeiträge zur quantitativen Steuerlehre*, 1–77.

Niemann, R. (2008). The effects of differential taxation on managerial effort and risk taking. *FinanzArchiv: Public Finance Analysis*, *64* (3), 273–310.

Niemann, R. (2011). Asymmetric taxation and performance-based incentive contracts. *CESifo Working Paper*, 1–25.

Niemann, R. & Simons, D. (2003). Costs, benefits, and tax-induced distortions of stock option plans. *Schmalenbach Business Review*, *55* (4), 321–341.

O'Byrne, S. F. (2014). Three Versions of Perfect Pay for Performance (Or The Rebirth of Partnership Concepts in Executive Pay). *Journal of Applied Corporate Finance*, *26* (1), 29–38.

Prendergast, C. (1999). The provision of incentives in firms. *Journal of Economic Literature*, *37* (1), 7–63.

Rappaport, A. (1986). *Creating shareholder value: the new standard for business performance*. New York: Free Press.

Sanders, W. G. & Tuschke, A. (2007). The adoption of institutionally contested organizational practices: The emergence of stock option pay in Germany. *Academy of Management Journal*, *50* (1), 33–56.

Schäfer, U. (2013). *Performance Measurement in langfristigen Prinzipal-Agenten-Beziehungen: Möglichkeiten und Grenzen einer Analyse auf Grundlage mehrperiodiger LEN-Modelle*. Baden-Baden: Nomos Verlagsgesellschaft mbH & Co. KG.

Schneider, D. (1977). Gewinnbesteuerung und Risikobereitschaft: zur Bewährung quantitativer Ansätze in der Entscheidungstheorie. *Zeitschrift für betriebswirtschaftliche Forschung*, *29*, 633–666.

Scholes, M. S., Wolfson, M. A., Erickson, M. M., Maydew, E. L. & Shevlin, T. (2009). *Taxes and business strategy: A planning approach* (4. Aufl.). Upper Saddle River, NJ: Pearson Prentice Hall.

Schöndube, J. R. (2006). *Nachverhandlungen in langfristigen Anreizbeziehungen* (1. Auflage Aufl.). DUV Deutscher Universitäts-Verlag.

Schreiber, M. (2018). *Milliarden-Boni, Mini-Dividende.* Zugriff am 31.10.2019, 12:30 Uhr auf https://www.sueddeutsche.de/wirtschaft/deutsche-bank-milliarden-boni-mini-dividende-1.3908840

Shleifer, A. & Vishny, R. W. (1997). A survey of corporate governance. *The Journal of Finance*, *52* (2), 737–783.

Spremann, K. (1987). Agent and principal. In G. Bamberg & K. Spremann (Hrsg.), *Agency theory, information, and incentives* (S. 3–37). Berlin: Springer.

Steck, A. (2019). *In den Grossbanken entsteht eine neue Bonus-Aristokratie.* Zugriff am 31.10.2019 auf https://nzzas.nzz.ch/wirtschaft/boni-fuer-banker-werden-trotz-kursverlust-ausgezahlt-ld.1467743?reduced=true

Stewart, G. B. (1991). *The quest for value: A Guide for Senior Managers*. United States of America: Harper Collins.

Voßmerbäumer, J. (2012). Effizienzwirkungen einer Regulierung von Managergehältern durch das Steuerrecht. *Schmalenbachs Zeitschrift für betriebswirtschaftliche Forschung*, *64* (5), 536–561.

Wagner, F. W. (1984). Grundfragen und Entwicklungstendenzen der betriebswirtschaftlichen Steuerplanung. *Betriebswirtschaftliche Forschung und Praxis*, *36* (3), 201–222.

Wagner, F. W. & Dirrigl, H. (1980). *Die Steuerplanung der Unternehmung.[Hauptbd.].* Fischer.

Wild, J. (1973). Organisation und Hierarchie. *Zeitschrift für Organisation*, *42* (1), 45–54.

Winter, S. (2001). Empirische Untersuchungen zur Managemententlohnung. In P.-J. Jost (Hrsg.), *Die prinzipal-agenten-theorie in der betriebswirtschaftslehre* (S. 491–539). Stuttgart: Schäffer-Poeschel.

Witt, P. (2001). Corporate Governance. In P.-J. Jost (Hrsg.), *Die prinzipal-agenten-theorie in der betriebswirtschaftslehre* (S. 85–115). Stuttgart: Schäffer-Poeschel.

Witzemann, T. & Currle, M. (2004). Controlling-Case Study - Bonusbanken - Unternehmenswertsteigerung und Managementvergütung langfristig verbinden. *Controlling*, *16* (11), 631–638.

Zwirner, C. & Tippelhofer, M. (2017/9). Rückstellungen für Compliance und Personal: Ansatz, Ausweis und Bewertung in Handels- und Steuerbilanz. In H. J. Voß & S. Stillers (Hrsg.), *Stub - unternehmensteuern und bilanzen* (Bd. 9, S. 1–16). Herne: Verlag Neue Wirtschaftsbriefe (NWB).

Rechtsquellenverzeichnis

Gesetze, Verordnungen und Verwaltungsanweisungen

[AktG]: Aktiengesetz vom 6. September 1965 (BGBl. I S. 1089), das zuletzt durch Artikel 9 des Gesetzes vom 17. Juli 2017 (BGBl. I S. 2446) geändert worden ist.

[AO]: Abgabenordnung in der Fassung der Bekanntmachung vom 1. Oktober 2002 (BGBl. I S. 3866; 2003 I S. 61), die zuletzt durch Artikel 10 des Gesetzes vom 11. Juli 2019 (BGBl. I S. 1066) geändert worden ist.

[EStG]: Einkommensteuergesetz in der Fassung der Bekanntmachung vom 8. Oktober 2009 (BGBl. I S. 3366, 3862), das zuletzt durch Artikel 1 des Gesetzes vom 25. März 2019 (BGBl. I S. 357) geändert worden ist.

[EStR]: Einkommensteuer-Richtlinien 2012 in der Fassung vom 16. Dezember 2005 (BStBl I Sondernummer 1/2005), geändert durch Einkommensteuer-Änderungsrichtlinien 2008 (EStÄR 2008) vom 18. Dezember 2008 (BStBl I S. 1017) und Einkommensteuer-Änderungsrichtlinien 2012 (EStÄR 2012) vom 25. März 2013 (BStBl I S. 276).

[GewStG]: Gewerbesteuergesetz in der Fassung der Bekanntmachung vom 15. Oktober 2002 (BGBl. I S. 4167), das zuletzt durch Artikel 8 des Gesetzes vom 11. Dezember 2018 (BGBl. I S. 2338) geändert worden ist.

[GmbHG]: Gesetz betreffend die Gesellschaften mit beschränkter Haftung in der im Bundesgesetzblatt Teil III, Gliederungsnummer 4123-1, veröffentlichten bereinigten Fassung, das zuletzt durch Artikel 10 des Gesetzes vom 17. Juli 2017 (BGBl. I S. 2446) geändert worden ist.

[HGB]: Handelsgesetzbuch in der im Bundesgesetzblatt Teil III, Gliederungsnummer 4100-1, veröffentlichten bereinigten Fassung, das zuletzt durch Artikel 3 des Gesetzes vom 10. Juli 2018 (BGBl. I S. 1102) geändert worden ist.

[KStG]: Körperschaftsteuergesetz in der Fassung der Bekanntmachung vom 15. Oktober 2002 (BGBl. I S. 4144), das zuletzt durch Artikel 2 des Gesetzes vom 4. August 2019 (BGBl. I S. 1122) geändert worden ist.

[SolZG]: Solidaritätszuschlaggesetz in der Fassung vom vom 15. Oktober 2002 (BGBl I S. 4131), das zuletzt durch Artikel 4 und 5 Gesetz zur steuerlichen Entlastung der Familien sowie zur Anpassung weiterer steuerlicher Regelungen (Familienentlastungsgesetz – FamEntlastG) vom 29. November 2018 (BGBl I S. 2210) geändert worden ist.

Urteile

BFH-Urteil vom 18.03.1965, IV 116/64 U, BStBl. III 1965, S. 289-291.

BFH-Urteil vom 14.05.1982, VI R 124/77, BStBl. II 1982, S. 469-474.

BFH-Urteil vom 01.08.1984, I R 88/80, BStBl. II 1985, S. 44-47.

BFH-Urteil vom 15.10.2003, XI R 17/02, BStBl. II 2004, S. 264-267.

BFH-Urteil vom 07.05.2015, VI R 44/13, BStBl. II 2015, S. 890-892.

FG Nürnberg-Urteil vom 10.05.2017, 3 K 1935/15, https://www.gesetze-bayern.de/Content/Document/Y-300-Z-BECKRS-B-2017-N-122815?hl=true, Abrufdatum: 14.10.2019.

Weitere Rechtsquellen

[BT-Drucksache 13/9712 (1998)]: Entwurf eines Gesetzes zur Kontrolle und Transparenz im Unternehmensbereich (KonTraG), Deutscher Bundestag, 13. Wahlperiode, S. 1-37.

[BT-Drucksache 16/13433 (2009)]: Beschlussempfehlung und Bericht des Rechtsauschusses (6.Ausschuss) zu dem Entwurf eines Gesetzes zur Angemessenheit der Vorstandsvergütung (VorstAG), Deutscher Bundestag, 16. Wahlperiode, S. 1-13.

[BT-Drucksache 19/9739 (2019)]: Entwurf eines Gesetzes zur Umsetzung der zweiten Aktionärsrechterichtlinie (ARUG II), Deutscher Bundestag, 19. Wahlperiode, S. 1-130.

[DCGK (2019)]: Deutscher Corporate Governance Kodex wie von der Regierungskommission am 9. Mai 2019 beschlossen, S. 1-34.

Junior Management Science 6(1) (2021) 149-189

Junior Management Science

journal homepage: www.jums.academy

Modern Approaches to Dynamic Portfolio Optimization

Philipp Schiele

Ludwig-Maximilians-Universität München

Abstract

Although appealing from a theoretical point of view, empirical assessments of dynamic portfolio optimizations in a mean-variance framework often fail to reach the high expectations set forth by analytical evaluations. A major reason for this shortfall is the imprecise estimation of asset moments and in particular, the expected return. This work levers recent advancements in the field of machine learning and employs three types of artificial neural networks in an attempt to improve the accuracy of the asset return estimation and the expected associated portfolio performances. After an introduction of the dynamic portfolio optimization framework and the artificial neural networks, their suitability for the considered application is analyzed in a two asset universe of a market and a risk-free asset. A comparison of the corresponding risk-return characteristics and those achieved using a more traditional exponentially weighted moving average estimator is subsequently drawn. While outperformance of the artificial neural networks is found for daily and monthly estimated returns, significance can only be established in the latter case, especially in light of trading costs. Multiple robustness checks are performed before an outlook for subsequent research opportunities is given.

Keywords: Portfolio optimization; machine learning; multilayer perceptron; convolutional neural network; long short-term memory.

1. Introduction

Since the establishment of Modern Portfolio Theory through the seminal work of Markowitz (1952) on portfolio selection, this field has seen a persistent rise of attention ever since. Indicated by both a vast body of literature as well as ever-growing interest from practitioners, the continued importance of ideal portfolio composition in the spirit of Markowitz is apparent, with the benefits of diversification and trade-offs between risk and return being cornerstones of modern investment practices.

A key implementational challenge to the originally proposed optimization problem remains as to how returns in financial markets can be forecast in order to make informed allocation decisions and avoid portfolio compositions that ex-post turn out to be suboptimal. Historically, the great difficulty of these forecasts has challenged both academia and practitioners alike. Generally, financial markets are known for a low signal-to-noise ratio, thereby making it hard to detect patterns within the data, to begin with. However, perhaps an even bigger challenge is posed by the constantly evolving and competitive nature of financial markets, which quickly incorporate the information contained in relationships having lead to an outperformance of some market participants, thereby changing the data generation process itself. As famously formulated in the random walk hypothesis (e.g. Cootner, 1964), many would even argue that forecasting financial returns is virtually impossible as prices are not predictable based on previous observations. This argument is closely linked to the efficient market hypothesis (e.g. Fama, 1970), a fundamental work of financial literature which in its strongest form states that markets immediately reflect all public and private information. Others, however, propose that markets are "efficiently inefficient" (e.g. Gârleanu & Pedersen, 2018), implying that at least minor information inefficiencies can persist, therefore potentially making prediction efforts worthwhile. Since Markowitz-style portfolio optimizations are highly sensitive to estimated future returns (Best & Grauer, 1991; Black & Litterman, 1992), uncovering even subtle and intricate patterns in financial data could indeed result in substantial differences in performance.

While the improvement of return forecasts is undoubtedly a major challenge, other fields have recently seen strong progress on tasks previously deemed hard by

DOI: https://doi.org/10.5282/jums/v6i1pp149-189

leveraging machine learning techniques. These include advances in medical applications like cancer detection (Esteva et al., 2017) or natural language processing (Radford et al., 2019). In particular, artificial neural networks seem to show continued improvements in model performance with increasing data sets (Sun, Shrivastava, Singh, & Gupta, 2017). While these models were already introduced by McCulloch and Pitts (1943), their recent resurgence is fueled by improved training algorithms and a sharp increase in computational power available to researchers (Livni, Shalev-Shwartz, & Shamir, 2014). Although in financial markets, the sheer amount of data generated on a daily basis is vast, the mentioned challenges unique to this field still have to be overcome in order to improve forecasts. Incidentally, the identification of complex patterns is one of the key strengths of artificial neural networks, which makes it natural to consider incorporating these techniques in the portfolio allocation context. The research objective is thus to investigate the usefulness of artificial neural networks in the context of dynamic portfolio optimization. More specifically, three types of artificial neural networks are being used to estimate the expected asset returns, followed by a subsequent assessment of the realized portfolio performances based on these estimates.

The remainder of the thesis is structured as follows: Chapter 2 reviews the literature concerning both dynamic portfolio optimization and artificial neural networks. Chapter 3 presents the methods and implementations used for the empirical analysis. In Chapter 4, the results of this empirical analysis are presented and discussed. Finally, Chapter 5 concludes the results of the thesis and presents an outlook for further research.

2. Literature Review

Since this thesis considers the use of neural networks in the context of dynamic portfolio optimization, both elements are introduced in this chapter. First, a brief overview of the considered portfolio optimization framework is given. Then, the different types of neural networks used for the succeeding analysis are reviewed, before the chapter concludes by formulating the research hypotheses.

2.1. Dynamic Portfolio Optimization

Modern Portfolio Theory is inevitably tied to Markowitz (1952), who introduced a framework considering a trade-off between the expected return of a portfolio and its risk. Markowitz measures the risk of a portfolio by its variance, and thus this approach is also referred to as the *mean-variance* framework. In this setting, a portfolio is considered efficient if it has the highest expected return for a given variance, or equivalently, the lowest variance for a given return. Formally, let $\mathbf{w}$ be the vector of asset weights and let $\boldsymbol{\mu}$ be the vector of their expected returns. Further, let $\boldsymbol{\Sigma}$ be the covariance matrix of these assets[1]. To find an efficient portfolio, one can minimize the portfolio variance σ_p^2 for a given expected portfolio return $\bar{\mu}$:

$$\min_{\mathbf{w}} \quad \sigma_p^2 = \mathbf{w}^\intercal \boldsymbol{\Sigma} \mathbf{w} \tag{1a}$$

$$\text{s.t.} \quad \boldsymbol{\mu}^\intercal \mathbf{w} \geq \bar{\mu} \tag{1b}$$

$$\mathbf{1}^\intercal \mathbf{w} = 1 \tag{1c}$$

$$(w_i \geq 0 \quad \forall i \in \{1, \ldots, n\}) \tag{1d}$$

This optimization problem can be stated as a quadratic program with linear constraints. The additional constraint 1c[2] ensures that the portfolio weights sum up to one. If the constraint 1d is added, short sales of assets are prevented. However, efficient portfolios can also be found by maximizing the expected portfolio return μ_p for a given level of portfolio variance $\bar{\sigma}$:

$$\begin{aligned} \max_{\mathbf{w}} \quad & \mu_p = \boldsymbol{\mu}^\intercal \mathbf{w} \\ \text{s.t.} \quad & \mathbf{w}^\intercal \boldsymbol{\Sigma} \mathbf{w} \leq \bar{\sigma} \\ & \mathbf{1}^\intercal \mathbf{w} = 1 \\ & (w_i \geq 0 \quad \forall i \in \{1, \ldots, n\}) \end{aligned} \tag{2}$$

Here, quadratically constrained linear programming can be used to solve the optimization problem. The set of efficient asset allocations that can be reached for varying $\bar{\sigma}$ or $\bar{\mu}$ values is referred to as the *efficient frontier*. Merton (1972) first suggested an analytical solution to finding these efficient portfolios in the case without the no short sales constraint 1d. However, additionally constrained settings often require an optimization for different risk or return targets to obtain the efficient frontier. The *optimal* portfolio is defined as the allocation for which an investors' indifference curve is tangential to the efficient frontier.

In this framework, only a single investment period is considered, making it *static*. However, it is of both theoretical and practical interest to optimize portfolios not only in a single period but *dynamically* in a multi-period setting, taking into account new information about the assets each period. Numerous contributions in this field have been made, such as Smith (1967), where at each point in time the portfolio is reevaluated based on the change in expected returns and covariances of the assets. The parameter estimation was performed by using a least-squares estimator, placing a higher weight on more recent data. The study considered a trade-off between the possibly inefficient allocation given the new moment estimates and costs involved when trading towards a new efficient allocation. The costs of trading thereby included

[1] Here and throughout the thesis, uppercase bold letters refer to matrices, lowercase bold letters to vectors and lowercase light letters to scalars.

[2] $\mathbf{1}$ represents the vector $(1, \ldots, 1)^\intercal$.

direct costs such as brokerage fees and indirect costs in the form of forgone profits on deferred tax obligations when taxable gains are realized early.

Brown and Smith (2011) start from a dynamic portfolio optimization model under the assumption of frictionless markets and consider different heuristic strategies to improve portfolio performances when this assumption is relaxed. These heuristics include the consideration of transaction costs in repeated static optimizations as well as the iterative maximization of expected utilities at a future date until which the chosen allocation needs to be held (i.e., no readjustment of the portfolio is possible). In numeric experiments, the performance of these heuristic strategies for different utility functions and levels of transaction costs is assessed, concluding that the strategies perform well in many cases.

A more analytical approach is taken by D. Li and Ng (2000), who derive an optimal investment policy for an unconstrained multiperiod mean-variance investor in the absence of trading costs. This provides a deepened understanding of portfolio optimality in a dynamic setting and can possibly be extended to include market frictions and portfolio constraints.

Other fundamental contributions to the general approach of optimizing dynamic investments include Mossin (1968), Samuelson (1969), Dumas and Luciano (1991) and Elton and Gruber (1974).

However, many of these theoretical contributions do not consider the question of how to optimally estimate moments of the asset return distribution, with the necessity of this estimation becoming clear when looking at the optimization problems themselves. Therefore, any empirical work on Modern Portfolio Theory needs to estimate the two main components of the optimization problems, namely the expected asset returns and their covariance matrix. Focusing on the estimation of the expected returns, Michaud (1989) states that in many applications the sample mean is used to estimate the expected asset returns, which he assesses to be a suboptimal choice, thus arguing for more advanced estimators. A crucial shortcoming of the sample mean in the estimation of time series moments is the equality of weights given to each observation, irrespective of their recency. Assuming more recent observations are more similar to future ones, the *exponentially weighted moving average (EWMA)* provides a natural extension to the sample mean by assigning exponentially decaying weights and thereby embedding the assumption into the estimation process (Severini, 2017). With this estimator, the expected return at time t can be recursively calculated as

$$\hat{\mu}_t = \lambda \hat{\mu}_{t-1} + (1-\lambda) r_t \tag{3}$$

for $t > 1$, where r_t refers to the return at time t, μ is the expected return and λ is a smoothing factor in the interval $(0, 1)$. Given its simplicity and the described inherent advantage over the sample mean, the EWMA estimator forms the baseline to which the neural networks are compared.

2.2. Neural Network Expected Return Estimators

Since the main research object is to evaluate the usefulness of artificial neural network expected return estimators in a portfolio optimization context, this section introduces the three different network architectures used throughout the analysis. In an abstract sense, artificial neural networks, henceforth just referred to as neural networks, can be thought of as versatile mappings of input and output values. This was formalized in the universal approximation theorem (Cybenko, 1989; Hornik, 1991) which states that finitely sized neural networks can approximate many continuous functions arbitrarily closely. In particular, these mappings can be nonlinear in their parameters, which makes them useful for many applications such as natural language processing (Goldberg, 2016), image processing (Egmont-Petersen, de Ridder, & Handels, 2002), and time series forecasting (Hill, O'Connor, & Remus, 1996). The latter scope makes them especially useful for the estimation of portfolio moments and provides the foundation for the considered analysis. Neural networks can be used in supervised and unsupervised learning settings. In supervised learning, the neural networks seek to minimize the prediction error made based on a model trained on data which contains the set of input values, so-called features, and the desired output values. In the context of financial forecasting, inputs could include historical returns, fundamental or macroeconomic data, with the outputs often represented by the future returns or just a binary representation of their sign. In contrast, unsupervised learning is concerned with tasks where no desired output is presented to the models explicitly. The focus of this thesis lies on supervised learning neural network applications, which can further be divided into regression and classification tasks, depending on whether class memberships or continuous outputs are modeled (Alpaydin, 2010). A general introduction to financial time series forecasting using neural networks is given by Kaastra and Boyd (1996) and Azoff (1994), with recent applications including the improvement of time series momentum strategies (Lim, Zohren, & Roberts, 2019), the forecast of ETF returns (Liew & Mayster, 2017) and the development of a statistical arbitrage strategy using neural network predictions (Krauss, Do, & Huck, 2017).

2.2.1. Multilayer Perceptrons

One widely used and sometimes referred to as the "vanilla" neural network architecture (Hastie, Tibshirani, & Friedman, 2009) is the *multilayer perceptron (MLP)*. As schematically represented in Figure 1, it consists of one input layer, one or more hidden layers, and one output layer. The total number of layers in the network is denoted L. More specifically, the illustrated network has

three input nodes, two hidden layers consisting of three nodes (or *units*) each and one output node. If two or more hidden layers are present, the network is referred to as a *deep* neural network (Bengio, 2009; LeCun, Bengio, & Hinton, 2015).

Following standard notation, the vector of input parameters is denoted as $\mathbf{x}$ and consists of the features $(x_1, \ldots, x_n)^\intercal$ and an additional bias term x_0 which can be seen as an intercept term, as it always takes on a value of one. For reasons of simplicity, this bias term is not depicted in the figures in this chapter. In the shown illustration, the input vector would thus be

$$\mathbf{x} = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{bmatrix} . \tag{4}$$

The values $\mathbf{a}^{(2)} = (a_1^{(2)}, \ldots, a_m^{(2)})^\intercal$ are obtained by first multiplying the input vector with the weight matrix $\mathbf{\Theta}^{(1)} \in \mathbb{R}^{m \times (n+1)}$ which assigns a single weight to each input value for every node in the first hidden layer. $\mathbf{\Theta}_{ij}^{(1)}$ thereby refers to the weight applied to the connection between the ith node in the first hidden layer and the jth node in the input layer. In general, when two adjacent layers are considered, let $n+1$ be the number of the preceding layers' nodes (including the bias unit) and let m be the number of nodes in the succeeding layer. Likewise, j always refers to a node in the preceding layer and i to a node in the succeeding layer. Again, in the illustrative example this weight matrix would be

$$\mathbf{\Theta^{(1)}} = \begin{bmatrix} \Theta_{10} & \Theta_{11} & \Theta_{12} & \Theta_{13} \\ \Theta_{20} & \Theta_{21} & \Theta_{22} & \Theta_{23} \\ \Theta_{30} & \Theta_{31} & \Theta_{32} & \Theta_{33} \end{bmatrix} . \tag{5}$$

This yields $\mathbf{z}^{(2)} = \mathbf{\Theta}^{(1)}\mathbf{x}$, a vector of weighted inputs to which an activation function $g(z)$ is element-wise applied. For this purpose, McCulloch and Pitts (1943) proposed a binary threshold function modelled after neuron activations in the brain. This function is defined as

$$g_0(z) = \begin{cases} 0 & \text{for } z \leq 0 \\ 1 & \text{for } z > 0 \end{cases} . \tag{6}$$

Since then, many activation functions have been suggested, such as the sigmoid function[3] (sometimes also referred to as the logistic or soft step function)

$$g_1(z) = \sigma(z) = \frac{1}{1+e^{-z}} , \tag{7}$$

the hyperbolic tangent function

$$g_2(z) = \tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} , \tag{8}$$

[3]Note that depending on the context, σ refers to the sigmoid activation function or the standard deviation.

or the rectified linear unit function (ReLU), introduced by Hahnloser, Sarpeshkar, Mahowald, Douglas, and Seung (2000) and again inspired by biological processes,

$$g_3(z) = \begin{cases} 0 & \text{for } z \leq 0 \\ z & \text{for } z > 0 \end{cases} = \max(z, 0) . \tag{9}$$

To add intuition on the functional principle of MLPs it should be noted that if the sigmoid function is used, each node in the hidden layer is a logistic regression on its input values, with the weights representing the regression coefficients. As in the MLP the activation of each node depends on all nodes in the previous layer, the hidden layers and the output layer are referred to as *fully connected* layers. As shown by Hayou, Doucet, and Rousseau (2018), the selection of an appropriate activation function needs to be performed under the consideration convergence performance, computational efficiency and the vanishing gradient problem, which is discussed below. Independent of the of activation function, the application of $g(z)$ on each element of $\mathbf{z}^{(2)}$ now yields $\mathbf{a}^{(2)}$. For each subsequent hidden layer k, again the previous nodes $\mathbf{a}^{(k-1)}$ and a bias term are multiplied by a weight matrix $\mathbf{\Theta}^{(k-1)}$ and the activation function is applied on the resulting vector $\mathbf{z}^{(k)}$ as shown in Equation 10:

$$\mathbf{a}^{(k)} = g(\mathbf{\Theta}^{(k-1)}\mathbf{a}^{(k-1)}) \tag{10}$$

For the output layer, $\mathbf{a}^{(L)}$, which again can consist of one or more nodes, the last hidden layer is multiplied by a final weight matrix $\mathbf{\Theta}^{(L-1)}$. Depending on the objective, different activation functions need to be applied here. For classification tasks, where output values in the interval $[0, 1]$ represent probabilities of a positive classification, often sigmoid or softmax functions are used, depending on whether exactly two or more than two classes are to be differentiated. For regression tasks where output values can lie outside of the interval $[0, 1]$, it is common to use no activation function (or equivalently, the linear activation function $g(z) = z$). Computing the output layer for a given input vector and weight matrices is referred to as *forward propagation* or a *forward pass*. As indicated by the arrows in Figure 1, the flow of information in an MLP is always directed from the input layer to the output layer. In particular, there are no cycles in this network architecture, which is why an MLP can be described as a *feedforward neural network*.

Backpropagation

Before a useful regression or classification can be performed on new input vectors using forward passes, the appropriate weight matrices have to be derived from training samples, which are sets of input vectors with known or *ground truth* output values denoted $\mathbf{y}$. The optimization of these weights is often achieved by employing the backpropagation algorithm, whose outline

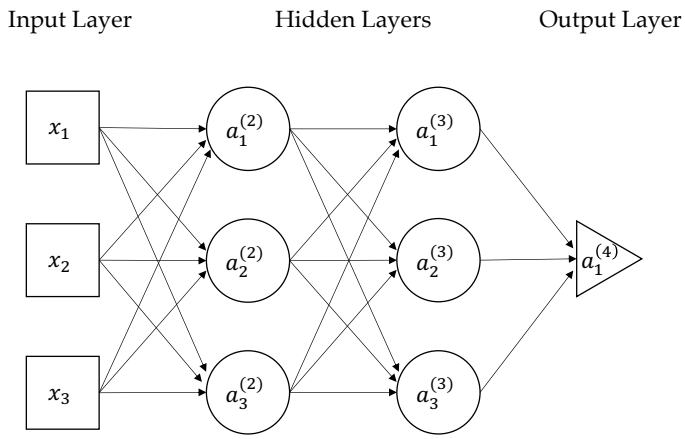


Figure 1: Schematic representation of an MLP

and origin is presented by Schmidhuber (2015). The underlying idea of backpropagation is to minimize the error made by the network on training data, starting with randomly initialized weights and iteratively adjusting them in the direction which decreases the overall error the most. A common analogy, among others used by F.-F. Li (2019), is to descend from a hill (point of high error) by only looking at the immediate surrounding and stepping in the direction which has the steepest downward slope (gradient) until the valley (a (local or global) minimum) is reached. For a single training example and randomly initialized weights, the activations of the output layer can be computed by performing a forward pass. Next, the error made using these weights in the output layer is calculated and denoted $J(\mathbf{\Theta})$. $\mathbf{\Theta}$ here refers to all individual weights from all layers, obtained by first unrolling all weight matrices $\mathbf{\Theta}^{(1)}, \ldots, \mathbf{\Theta}^{(L-1)}$ into vectors and concatenating them. A typical error function for regressions is the *mean squared error function* (*MSE*, Equation 11) with the *categorical cross-entropy function* (*CCE*, Equation 12) often being used for classification tasks.

$$J_{MSE}(\mathbf{\Theta}) = \frac{1}{m} \sum_{\nu=1}^{m} (a_\nu^{(L)} - y_\nu)^2 \tag{11}$$

$$J_{CCE}(\mathbf{\Theta}) = -\sum_{\nu=1}^{m} y_\nu \log(a_\nu^{(L)}) \tag{12}$$

The derivative of this error with respect to $\mathbf{\Theta}$ is indicative of the necessary adjustments that need to be made in order to decrease the error. It is also referred to as the gradient of $\mathbf{J}$ and can be written as

$$\nabla \mathbf{J} = \frac{\partial J(\mathbf{\Theta})}{\partial \mathbf{\Theta}} = \begin{bmatrix} \frac{\partial J(\mathbf{\Theta})}{\partial \Theta_{10}^{(1)}} \\ \vdots \\ \frac{\partial J(\mathbf{\Theta})}{\partial \Theta_{m(n+1)}^{(L-1)}} \end{bmatrix}. \tag{13}$$

For each of the weights in the final layer, $\Theta_{ij}^{(L-1)}$, one can apply the chain rule as shown in Equation 14 in order to decompose the derivative of the error with respect to this weight into a change in $z_i^{(l)}$ caused by the initial change in $\Theta_{ij}^{(L-1)}$. This in turn changes $a_i^{(L)}$ through the activation function, thereby directly influencing the error term $J(\mathbf{\Theta})$.

$$\frac{\partial J(\mathbf{\Theta})}{\partial \Theta_{ij}^{(L-1)}} = \frac{\partial J(\mathbf{\Theta})}{\partial z_i^{(L)}} \frac{\partial z_i^{(L)}}{\partial \Theta_{ij}^{(L-1)}} = \frac{\partial J(\mathbf{\Theta})}{\partial a_i^{(L)}} \frac{\partial a_i^{(L)}}{\partial z_i^{(L)}} \frac{\partial z_i^{(L)}}{\partial \Theta_{ij}^{(L-1)}} \tag{14}$$

Next, the derivatives of $J(\mathbf{\Theta})$ with respect to the weights in the penultimate layer, $\mathbf{\Theta}^{(L-2)}$, are needed. Again, using the chain rule, this derivative can be decomposed as follows:

$$\frac{\partial J(\mathbf{\Theta})}{\partial \Theta_{ij}^{(L-2)}} = \frac{\partial J(\mathbf{\Theta})}{\partial z_i^{(L-1)}} \frac{\partial z_i^{(L-1)}}{\partial \Theta_{ij}^{(L-2)}} = \frac{\partial J(\mathbf{\Theta})}{\partial a_i^{(L-1)}} \frac{\partial a_i^{(L-1)}}{\partial z_i^{(L-1)}} \frac{\partial z_i^{(L-1)}}{\partial \Theta_{ij}^{(L-2)}} \tag{15}$$

However, $\frac{\partial J(\mathbf{\Theta})}{\partial a_i^{(L-1)}}$ is not known but needs to be derived first. The node $a_i^{(L-1)}$ contributes to the total error through all its connections to the output layer. Since the layers are fully connected, the error induced in all

output nodes needs to be considered if $a_i^{(L-1)}$ changes. A change in in $a_i^{(L-1)}$ introduces a change in $\mathbf{z}^{(L)}$ which again through the activation function changes $\mathbf{a}^{(L)}$, leading to a change in $J(\mathbf{\Theta})$. Thus, it follows that:

$$\frac{\partial J(\mathbf{\Theta})}{\partial a_j^{(L-1)}} = \sum_{i=1}^{m} \frac{\partial z_i^{(L)}}{\partial a_j^{(L-1)}} \frac{\partial a_i^{(L)}}{\partial z_i^{(L)}} \frac{\partial J(\mathbf{\Theta})}{\partial a_i^{(L)}} \tag{16}$$

Note that the index of $a^{(L-1)}$ was changed from i to j as now not the layers $L-2$ and $L-1$, but $L-1$ and L are considered. Also note that the derivative with respect to $a_0^{(L-1)}$, the bias node, is omitted since its activation is fixed at a value of 1. With this, the derivative of the error with respect to the weights of all layers can be obtained. This is achieved by propagating the error from the output layer back to the currently analyzed layer through a recursive application of the single-layer error propagation introduced in Equation 16. This behavior is eponymous for the algorithm. There, the error can then be decomposed similarly to Equation 15. A generalized form of the derivative of the loss functions with respect to the weights for all previous layers $l = L-3, ..., 1$ is shown in Equation 17 which makes use of recursively propagating the error from the output layer to layer l by making use of Equation 18.

$$\frac{\partial J(\mathbf{\Theta})}{\partial \Theta_{ij}^{(l)}} = \frac{\partial J(\mathbf{\Theta})}{\partial z_i^{(l+1)}} \frac{\partial z_i^{(l+1)}}{\partial \Theta_{ij}^{(l)}} = \frac{\partial J(\mathbf{\Theta})}{\partial a_i^{(l+1)}} \frac{\partial a_i^{(l+1)}}{\partial z_i^{(l+1)}} \frac{\partial z_i^{(l+1)}}{\partial \Theta_{ij}^{(l)}} \tag{17}$$

$$\frac{\partial J(\mathbf{\Theta})}{\partial a_j^{(l+1)}} = \sum_{i=1}^{m} \frac{\partial z_i^{(l+2)}}{\partial a_j^{(l+1)}} \frac{\partial a_i^{(l+2)}}{\partial z_i^{(l+2)}} \frac{\partial J(\mathbf{\Theta})}{\partial a_i^{(l+2)}} \tag{18}$$

At this point, the complete gradient vector $\nabla\mathbf{J}$ is known and thus the weight vector $\mathbf{\Theta}$ can be adjusted to reduce the error. This is done by an incremental change in the negative direction of the gradient. The improved weight vector, denoted $\mathbf{\Theta}^+$, is thus given by:

$$\mathbf{\Theta}^+ = \mathbf{\Theta} - \alpha\nabla\mathbf{J} \tag{19}$$

An iterative application of forward propagations and backpropagations is used to further improve weights until ideally the gradient vector approaches $\mathbf{0}$ and a minimum of the error function is reached. The constant $\alpha > 0$ is referred to as the *learning rate* and has to be chosen based on a trade-off between convergence speed and convergence probability (Bengio, 2012). In order to simplify the notation, the description of the backpropagation algorithm above only focuses on one training example. For a set of o training samples, $\{(x^{(1)}, y^{(1)}, ..., x^{(o)}, y^{(o)}\}$, the overall gradient $\nabla\mathbf{J}_{total}$ is defined as the arithmetic mean of all individual sample gradients:

$$\nabla\mathbf{J}_{total} = \frac{1}{o}\sum_{p=1}^{o} \nabla\mathbf{J}_p \tag{20}$$

Gradient Descent

Starting from a random initialization and repeatedly updating the weights using the overall gradient $\nabla\mathbf{J}_{total}$ until convergence is reached is referred to as *(batch) gradient descent (GD)*. Algorithm 1 gives an overview of the gradient descent algorithm, making use of both forward propagation and backpropagation.

```
Data: Training samples {(x^(1), y^(1), ..., x^(o), y^(o))}
Randomly initialize Θ
while not converged do
    Set ∇J = 0
    for i = 1 to o do
        Set ∇J_i = 0
        Set a^(1) = x^(i)
        Perform forward propagation to compute
          a^(l) for l = 2, 3, ..., L
        for l = L-1 to 1 do
            Obtain ∂J(Θ)/∂a^(l+1) by backpropagation the
              error
            Compute ∂J(Θ)/∂Θ^(l) and concatenate to ∇J_i
        end
        ∇J = ∇J + (1/o) ∇J_i
    end
    Θ = Θ − α∇J
end
```

Algorithm 1: Gradient descent

An alternative implementation exists where the weights are not updated based on the cumulative gradient of all samples, but rather after each individual sample based on the gradient as measured by this sample alone. This method is called *stochastic gradient descent (SGD)*. The name refers to the stochastic approximation of the overall gradient descent algorithm through repeated adjustments based on the single gradients. This approximation is often supported by shuffling the data set first in order to avoid clustering. There are several advantages and disadvantages to both GD and SGD, such as the stable gradient of GD and more memory efficient implementations of SGD (since only one sample needs to be held in memory at any point in time, not the complete data set). Thus, for practical applications oftentimes *mini-batch gradient descent* is used, which aims to combine the advantages of GD and SGD without introducing too many of the disadvantages. In mini-batch gradient descent, the weights are updated after the gradient of a subset of n training examples is computed, with $1 < n < o$. Its main advantages are a more stable gradient compared to SGD while still, only the samples of the current batch need to be held in memory. On the other hand, one drawback is that the batch size is an additional variable which needs to be chosen by the researcher. An extended overview of the introduced gradient descent variations is given by Ng

(2019a).

After choosing a subtype of the gradient descent algorithm, the weights of the MLP network can be optimized using the training data and predictions can subsequently be made using forward passes on new samples, thereby completing the functional principle of the MLP.

2.2.2. Convolutional Neural Networks

Another class of neural networks is the *convolutional neural network (CNN)*. In their current form, CNNs were pioneered by LeCun et al. (1989, 1990), with a comprehensive introduction being provided by Lecun, Bottou, Bengio, and Haffner (1998). Until today, following their original domain of application, CNNs are most commonly used in the analysis of image data (Ciresan, Meier, & Schmidhuber, 2012; Lawrence, Giles, Ah Chung Tsoi, & Back, 1997). However, their applications also extend beyond image data to tasks like natural language processing (Collobert et al., 2011; Grefenstette, Blunsom, de Freitas, & Hermann, 2014; Kalchbrenner, Grefenstette, & Blunsom, 2014; Kim, 2014) and even time series forecasting (Borovykh, Bohte, & Oosterlee, 2017). As shown in Figure 2, CNNs differ from MLPs through the inclusion of oftentimes multiple and alternating *convolutional layers* and *pooling layers* and usually a single *flatten layer*, which is further discussed below. In univariate time series applications, the input vector $\mathbf{x}$ represents the individual observations of the time series. Multivariate time series or gray-scale image data, on the other hand, require the input layer to be a two-dimensional matrix, with colored images or video data requiring even higher dimensional inputs.

Convolutional Layers

The convolutional layers of a CNN are at the core of its functional principle. As shown in Figure 2, these layers take subsets of the input layer and apply a *filter* (also called *kernel*) to them, resulting in a weighted sum of these input values. This aggregation process is referred to as a *convolution* and is intended to extract higher-level features from the input while retaining spatial structures. For the more general case with a two dimensional input, let this input layer be a matrix $\mathbf{X}$ with dimensions $[k \times l]$. Further, let the kernel be the weight matrix $\mathbf{\Theta}^{(1)}$ with dimensions $[m \times n]$. Convolutions usually are performed on a central element and its surroundings, thus m and n are uneven in these cases. In the one-dimensional example depicted in Figure 2, the dimensions of the input are $[8 \times 1]$ with a kernel size of $[3 \times 1]$. The kernel is thus given by:

$$\mathbf{\Theta}^{(1)} = \begin{bmatrix} \Theta_{11} \\ \Theta_{21} \\ \Theta_{31} \end{bmatrix} \tag{21}$$

Define $\mathbf{X}_{sub}$ as a subset of $\mathbf{X}$ with the shape of the kernel $[m \times n]$. For this subset, the scalar output of a convolution is given by

$$z_{k_i l_j} = \sum_{\nu=1}^{m} \sum_{\eta=1}^{n} \Theta^{(1)}_{\nu\eta} X_{sub,\nu\eta} = \mathbf{1}^{\intercal}(\mathbf{\Theta}^{(1)} \odot \mathbf{X}_{sub}))\mathbf{1} \tag{22}$$

with $\odot$ denoting the Hadamard product. A convolution, therefore, is the sum of the elements resulting from an element-wise multiplication of $\mathbf{X}_{sub}$ and the kernel $\mathbf{\Theta}^{(1)}$[4]. The indices of the output $z_{k_i l_j}$ thereby refer to the position of the central element of $\mathbf{X}_{sub}$ within $\mathbf{X}$. While theoretically conceivable, it is uncommon to apply a nonlinear activation function to this output, thus setting $a_{k_i l_j} = z_{k_i l_j}$. The full convolutional layer is obtained by repeating the convolution process for each possible subset $\mathbf{X}_{sub}$ in $\mathbf{X}$. More specifically, starting in one corner of $\mathbf{X}$, after each convolution the filter is offset by s in the horizontal direction, where s is referred to as the *stride* parameter. This is repeated until no further horizontal shifting is possible. Then, the kernel is reset to its original horizontal position and shifted by s in the vertical direction before starting another horizontal pass. This is repeated until the end of the input matrix is reached. This "serpentine" shifting pattern of the kernel can also be visualized as a typewriter which moves from left to right until the end of the line is reached and it starts again from the left in the next line. It is important to note here that in contrast to a fully connected layer the spatial structure in the input layer is contained during the convolution since one node in the output layer only contains information from the surrounding nodes in the input layer. More specifically, the same weights are used whenever the filter is moved across the input, which is referred to as *weights sharing*. In the one-dimensional case, a horizontal shift is not possible, so only vertical shifts take place. Thus, as shown in Figure 2, for a stride value of $s = 1$, after the initial convolution, the filter can be shifted five times vertically. The size of the convolutional layer is therefore given as $[6 \times 1]$. In general, in a one-dimensional setting the length of the convolutional layer is given by

$$k_a = \left\lfloor \frac{k-m}{s} \right\rfloor + 1\,. \tag{23}$$

However, for some applications, it is desirable to have convolutional layers with the same shape as the input layer. In these cases, the input matrix can be extended by padding inputs with a value of zero (Goodfellow, Bengio, & Courville, 2016). As again shown in Figure 2, different kernels can be used in order to extract different higher-level features from the input, thus resulting in another dimension for the output of the convolutional layer.

[4] In some applications the kernel is flipped both horizontally and vertically to differentiate a convolution from a cross-correlation. While results would vary for a given, non-symmetric kernel, this has no effect in a CNN since the filter weights are not given but learned by the model.

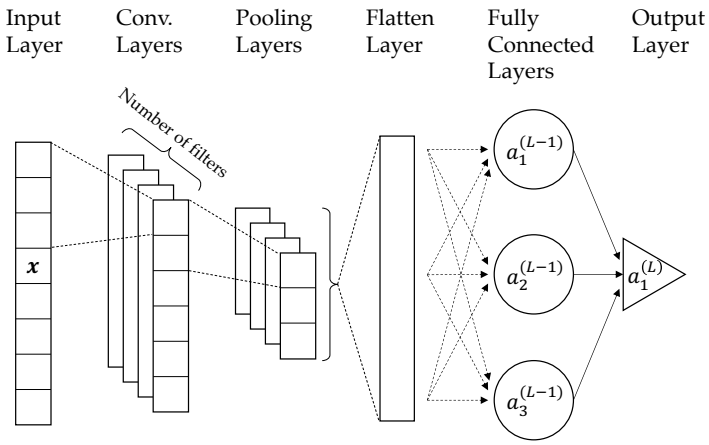


Figure 2: Schematic representation of a CNN

Pooling Layers
After a convolutional layer, often a pooling layer is used in order to reduce the dimensionality of the layers. Since CNNs often have a large input matrix, this reduction in dimensionality can lead to a significant improvement in the training speed of the model and also reduces its number of trainable parameters. The underlying idea is to aggregate neighboring nodes, ideally capturing most of their information content while reducing the layer dimensions. Similar to the convolutions, an aggregation function is applied to subsets of nodes of the previous layer with size $[m \times n]$, where m and n can be different from the filter size of previous layers. Popular aggregations include the arithmetic mean (*average pooling*) or the maximum (*max pooling*) of these nodes (Goodfellow et al., 2016). Again, this subset is moved along the previous layer by stride increments of s, which is often chosen such that the subsets do not overlap. In Figure 2 the aggregation subsets are chosen to have size $[2 \times 1]$ with $s = 2$. This reduces the size of the layers by 50% from $[6 \times 1]$ to $[3 \times 1]$.

Flatten Layer
In a CNN the number of filters used determines the number of resulting convolutional layers. Turning again to Figure 2, it is apparent that the number of layers also persists when pooling layers are applied. In order to use the higher-level features extracted through convolutions and pooling, a single matrix needs to be formed based on the forwardmost convolutional or pooling layer. This is performed by a flatten layer, which effectively concatenates all matrices of the preceding layer.

Fully Connected Layers
The fully connected layers can finally take the extracted higher-level features from the previous layers and model their dependencies, just like a discrete MLP network would if these features were provided externally. This behavior makes CNNs especially useful for extensive data sets with many potential features, since learning these features is part of the model training, thus requiring little domain knowledge for manual feature engineering (Kang, Ye, Li, & Doermann, 2014). Rather, the network itself can first find higher-level features and then train an MLP network on them.

Gradient Descent
Like MLP networks, CNNs are often trained from a random initialization using the gradient descent algorithm. Thus, after a forward pass, as described above, was performed, the error made using these weights is computed. For the fully connected layers, the derivative of the error with respect to their weights can again be obtained through the application of Equation 17. Arriving at the flatten layer, no derivative with respect to weights needs to be computed since this layer is merely a reshaping of the preceding nodes and it has no impact on the gradient. Since pooling layers also do not have weights which are learned by the model, again no derivative with respect to their weights can be computed. They do, however, affect the composition of the error through the way they aggregate the nodes of the previous layer. Thus, when average pooling is used, the error also has to be equally split to the nodes of the previous layer. For a subset size of $[m \times n]$, the error is thus multiplied by $\frac{1}{mn}$ when passing it to the previous nodes. For max pooling layers, the full error is propagated to the "winning" node of the forward pass (i.e., the node with the maximum activation, the value of which was used by the max pooling node) and an error of 0 is propagated to all other nodes, since their activations eventually had no effect on the error. Next, the derivative

of the error with respect to the weights of a convolutional layer is considered. Since a filter shares the same weights when moved across the input layer, it is intuitive to see that a single weight in a filter contributes to the total error through multiple paths. Thus, the total contribution of a weight in a filter is given by the sum of the derivatives of the backpropagated errors over all nodes of the convolutional layers' output nodes. Since convolutional layers are not restricted to succeed the input layer, the output of the convolutional layer is denoted $\mathbf{a}^{(l)}$ and its inputs $\mathbf{a}^{(l-1)}$. Further, let $[k^{(l-1)} \times l^{(l-1)}]$ denote the dimensions of $\mathbf{a}^{(l-1)}$, with $[k^{(l)} \times l^{(l)}]$ denoting the dimensions of $\mathbf{a}^{(l)}$. Then, the total contribution of weight $\Theta_{ij}^{(l-1)}$ is given by

$$\frac{\partial J(\mathbf{\Theta})}{\partial \Theta_{ij}^{(l-1)}} = \sum_{\nu=1}^{k^{(l)}} \sum_{\eta=1}^{l^{(l)}} \frac{\partial J(\mathbf{\Theta})}{\partial a_{\nu\eta}^{(l)}} \frac{\partial a_{\nu\eta}^{(l)}}{\partial z_{\nu\eta}^{(l)}} \frac{\partial z_{\nu\eta}^{(l)}}{\partial \Theta_{ij}^{(l-1)}} . \tag{24}$$

In case there are other layers preceding the convolutional layer, also the derivative of the error with respect to the input layer $\mathbf{a}^{(l-1)}$ needs to be obtained. Again, each node in this layer can contribute to the overall error through multiple paths since moving a filter across the input layer exposes the node to different weights in the filter matrix. The total contribution of node $a_{ij}^{(l-1)}$ is again the sum over all these paths through the different weights of the filter, as shown in Equation 25.

$$\frac{\partial J(\mathbf{\Theta})}{\partial a_{ij}^{(l-1)}} = \sum_{\nu=1}^{k^{(l)}} \sum_{\eta=1}^{l^{(l)}} \frac{\partial z_{\nu\eta}^{(l)}}{\partial a_{ij}^{(l-1)}} \frac{\partial a_{\nu\eta}^{(l)}}{\partial z_{\nu\eta}^{(l)}} \frac{\partial J(\mathbf{\Theta})}{\partial a_{\nu\eta}^{(l)}} \tag{25}$$

Since convolutional layers are not fully connected, the sparsity of the connections results in $\frac{\partial z_{\nu\eta}^{(l)}}{\partial a_{ij}^{(l-1)}} = 0$ for unconnected nodes.

With that, the error can be propagated back through pooling and convolutional layers, and thus the gradient vector $\nabla \mathbf{J}$ can be obtained. The gradient descent algorithm can then be applied, as shown in Algorithm 1.

2.2.3. Recurrent Neural Networks

While MLPs and CNNs evaluate each sample of inputs and outputs independently from others, for specific tasks, it would be beneficial to model a dependence structure between samples. In fact, sequential tasks like the modeling of the next word based on parts of a sentence or a time series prediction often require information about multiple preceding samples. One approach would be to feed multiple previous values to an MLP network. However, one could also conceptualize to only feed one sample at a time to a network which itself keeps information about previous samples. This is where *recurrent neural networks (RNNs)* emerge. As insinuated in Figure 3, at every step t in the training process, the inputs of an RNN do encompass not only the original input features $\mathbf{x}_t$ at that point, but also information about previous samples.

Elman Networks

Like for most neural network architectures, many implementations of RNNs exist. The specific network depicted in Figure 3 and described in the following was suggested by Elman (1990) and is thus referred to as the *Elman network*. In this network, information is passed from $t-1$ to t by providing the activations of the hidden layer as further inputs to the network.

Mathematically, at each time step t, this is achieved by first multiplying a weight matrix $\mathbf{\Theta}_1$ by the input vector $\mathbf{x}_t$. Next, the weight matrix $\mathbf{\Theta}_2$ is multiplied by the previous samples' hidden layer activations $\mathbf{a}_{t-1}^{(L-1)}$. Both results are then added element-wise to obtain the inputs for the hidden layer, $\mathbf{z}_t^{(L-1)}$. Like in any neural network, in order to obtain the hidden state, an activation function $g(z)$ is applied to each element of $\mathbf{z}_t^{(L-1)}$, as shown in Equation 26. While the dimensions of $\mathbf{x}_t$ and $\mathbf{a}_{t-1}^{(L-1)}$, i.e., the input vector and previous activations, can differ, it is required for $\mathbf{\Theta}_1$ and $\mathbf{\Theta}_2$ to match in their first dimension such that the resulting vectors can be added element-wise.

Based on the hidden layer $\mathbf{a}_t^{(L-1)}$, the output layer $\mathbf{a}_t^{(L)}$ can again be computed similar to the MLP case, namely by multiplying the weight matrix $\mathbf{\Theta}_3$ by the hidden activations and applying an activation function.

$$\mathbf{a}_t^{(L-1)} = g(\mathbf{\Theta}_1 \mathbf{x}_t + \mathbf{\Theta}_2 \mathbf{a}_{t-1}^{(L-1)}) \tag{26}$$

$$\mathbf{a}_t^{(L)} = g(\mathbf{\Theta}_3 \mathbf{a}_t^{(L-1)}) \tag{27}$$

For the next time step $t+1$, the hidden layer of t, $\mathbf{a}_t^{(L-1)}$, is combined with a new input vector $\mathbf{x}_{t+1}$ and with the same weights $\mathbf{\Theta}_1$, $\mathbf{\Theta}_2$ and $\mathbf{\Theta}_3$ again the hidden and output layer can be computed. This process is repeated iteratively until the end of the sequence is reached.

In order to keep Figure 3 easily understandable, vectors, matrices and single connections between them were drawn. However, just like in the MLP case in Figure 1, each line is fully connecting all nodes of one layer to all nodes of the subsequent layer. Further, it should be noted that the activation functions $g(z)$ of the hidden and output layer do not necessarily have to be identical. This is especially relevant as the activation function of the output layer often is determined by the application of the neural network.

Considered at each point in time, simple RNNs usually consist of only one hidden layer. The *deepness* of this network architecture rather comes from the possibility to access information from previous observations. This is achieved by first performing a forward pass as described above, starting at the first time step. Since for the first time step, no previous activations are present, a randomly initialized vector or a vector of zeros can be used. This

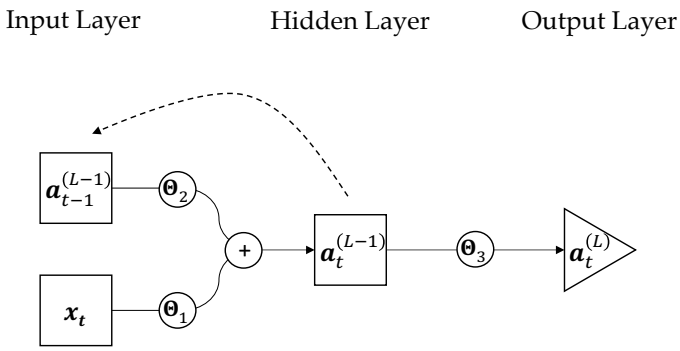


Figure 3: Schematic representation of an RNN

forward pass can be conceptualized by chaining copies of Figure 3, only adjusting the time index at each point. This is also called unfolding the network (Goodfellow et al., 2016). Then, the error at each time step can be computed, resulting in a loss of $J(\mathbf{\Theta})$. Through propagating the error back from the final to the first observation and applying the chain rule as shown in Equation 17, $\frac{\partial J(\mathbf{\Theta})}{\mathbf{\Theta}_1}$, $\frac{\partial J(\mathbf{\Theta})}{\mathbf{\Theta}_2}$ and $\frac{\partial J(\mathbf{\Theta})}{\mathbf{\Theta}_3}$ are obtained. The weight matrices are then adjusted before another pass is started. Since this retains the time dimension of the observations, this method of training is referred to as *backpropagation through time (BPTT)*, which was proposed by Werbos (1988), Mozer (1989) as well as Robinson and Fallside (1987).

Another popular simple RNN network was introduced by Jordan (1997), with the major differentiating factor being that not the previous hidden layer $\mathbf{a}_{t-1}^{(L-1)}$, but the previous output layer $\mathbf{a}_{t-1}^{(L)}$ is provided as an additional input at time t.

Long Short-Term Memory Networks

RNNs allow for information to persist in a network between observations. However, if the relevant information was not contained in a recent observation, but rather many observations before, simple RNNs can have great difficulty to represent these dependencies. Although there is no theoretical limit to how far back the information in these RNNs can persist, in empirical applications it has been difficult to model long term dependencies with them (Bengio, Simard, & Frasconi, 1994). The so-called *vanishing gradient problem*, as analyzed by Hochreiter (1991) and Hochreiter, Bengio, Frasconi, and Schmidhuber (2001), was found to play a major role in this restriction. Vanishing gradients describe a phenomenon where the gradient continuously decreases with the distance to the output layer. One example of this phenomenon can be considered where the chain rule is applied repeatedly by the backpropagation algorithm, as described in Equation 18. The derivative of many common activation functions, $\frac{\partial a}{\partial z} = g'(z)$, is bound to the range $(0, 1)$. The derivative of the popular sigmoid function (7) is even bound to the range $(0, \frac{1}{4})$. Thus, when these activation functions are used, and the error is propagated back many layers, the multiplicative linkage of these derivatives tends towards zero exponentially. One possible mitigation can be to use activation functions which do not exhibit this property, such as the ReLU function (9). However, while here the derivative equals one for $z > 0$, a single node with $z \leq 0$ leads to multiplication with zero and thus stops the gradient for earlier layers. Another approach to counteract the vanishing gradient problem is to use network architectures based on RNNs which are specifically designed to pass information along for many time steps, such as the *Long Short-Term Memory (LSTM)* network. Building on previous work in the field of RNNs, the LSTM architecture was introduced by Hochreiter and Schmidhuber (1997). It was much more recent, however, that using larger data sets on this model has produced numerous state-of-the-art results in speech recognition (Graves, Mohamed, & Hinton, 2013), handwriting recognition (Graves et al., 2009), the supersedence of humans in complex computer games (OpenAI, 2018) and also time series forecasting (Laptev, Yosinski, Li, & Smyl, 2017).

Essential to the ability of these networks to retain information over extended periods is the so-called cell state $\mathbf{c}$, also called the memory of the model. This vector can hold information for many time steps, or more specifically until it is *actively* forgotten. A full overview of the network architecture at time t is again given in Figure 4.

Looking at the functional principles of the network, there are four layers with individual weights in an LSTM which control the flow of information, indicated by the ellipses in the figure. These layers, also referred to as *gates*, are defined as follows:

The *forget gate* (Equation 28), as implied by the name, is trained to forget information by selectively removing it from the cell state, based on the current inputs $\mathbf{x}_t$ and

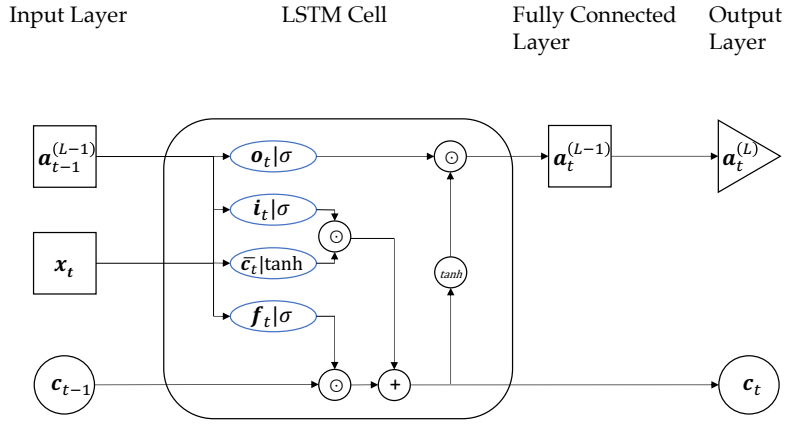


Figure 4: Schematic representation of an LSTM Network

$\mathbf{a}_{t-1}^{(L-1)}$. This is achieved by applying the sigmoid function to the weighted inputs, thus mapping them into the interval $(0,1)$. Since the Hadamard product of the resulting vector and the cell state is computed, values close to one correspond to a retention of the information in the cell state, whereas values close to zero discard most of the information.

$$\mathbf{f}_t = \sigma(\mathbf{x}_t * \Theta_{x,f} + \mathbf{a}_t^{(L-1)} * \Theta_{a,f}) \tag{28}$$

$$\bar{\mathbf{c}}_t = \tanh(\mathbf{x}_t * \Theta_{x,c} + \mathbf{a}_t^{(L-1)} * \Theta_{a,c}) \tag{29}$$

$$\mathbf{i}_t = \sigma(\mathbf{x}_t * \Theta_{x,i} + \mathbf{a}_t^{(L-1)} * \Theta_{a,i}) \tag{30}$$

$$\mathbf{o}_t = \sigma(\mathbf{x}_t * \Theta_{x,o} + \mathbf{a}_t^{(L-1)} * \Theta_{a,o}) \tag{31}$$

Similarly, information can also be added to the cell state dependent on the inputs through a combination of the *candidate gate* (Equation 29) and the *input gate* (Equation 30). The candidate gate first produces possible values to update the cell state with. Since it has a hyperbolic tangent activation function, the results lie in the interval $(-1,1)$ such that both positive and negative states can be achieved. The input gate, on the other hand, has a sigmoid activation function in order to control the extent to which each of the candidate nodes is added to the cell state. The described updating process of the cell state is expressed in Equation 32. With the cell state updated, the output of the network can now be computed. Therefore, first the *output cell* (Equation 31) combines the weighted inputs $\mathbf{x}_t$ and $\mathbf{a}_{t-1}^{(L-1)}$ and applies a sigmoid activation function. These activations are now amended by the information of the cell state. In order to make sure the cell state is in the interval $(-1,1)$, the hyperbolic tangent function is applied to it again. Finally, the Hadamard product of the output gate and the squashed cell state yield the hidden layer output of the LSTM cell, $\mathbf{a}_t^{(L-1)}$, as shown in Equation 33. The multiplicative link of the cell state thus can either weaken or even invert the activation of the output gate. The effect of this step can be conceptualized well with an example from natural language modeling. A negating word at the beginning of a sentence might be held in the cell state with a negative value, which would then invert the prediction otherwise made based on words occurring much later in the sentence.

$$\mathbf{c}_t = \mathbf{c}_{t-1} \odot \mathbf{f}_t + \mathbf{i}_t \odot \bar{\mathbf{c}}_t \tag{32}$$

$$\mathbf{a}_t^{(L-1)} = \mathbf{o}_t \odot \tanh(\mathbf{c}_t) \tag{33}$$

To obtain the appropriate activation for the task at hand, a final fully connected layer is added before reaching the output layer, which yields $\mathbf{a}_t^{(L)}$. Although the LSTM is is more complex than a simple RNN, it is also trained using the BPTT algorithm by first unrolling the network and propagating the error back from the last to the first observation, finally updating the weight matrices $\Theta_{x,\nu}$ and $\Theta_{a,\nu}$ $\forall\, \nu \in \{f,c,i,o\}$.

Again, numerous adaptions to the LSTM model were suggested since its introduction. Gers and Schmidhuber (2000) introduce so-called *peepholes* to the network, which add the incoming cell state $\mathbf{c}_{t-1}$ as inputs to the gates. Another alteration was suggested by Cho et al. (2014), who combine the functionally related candidate and input gates to a single gate referred to as the *update gate*.

2.3. Hypotheses formulation

Following the presented literature, the research objective is to compare the performance of portfolios optimized using expected return estimates stemming from the introduced neural networks to those optimized using a more traditional EWMA estimator. It is well known

that in the presented framework the estimation of the expected return is a key determinant of the resulting portfolio performance as the optimal portfolio weights are highly sensitive to these estimates (Best & Grauer, 1991; Black & Litterman, 1992). If neural networks indeed are superior in estimating the expected return, a substantial increase in portfolio performance is to be expected. As higher portfolio returns can also come at the expense of increased volatility, the Sharpe ratio (Sharpe, 1966) of the portfolios is chosen as a performance metric to address this trade-off.

With the presented EWMA estimator referred to as the *baseline estimator*, and the MLP, CNN and LSTM estimators jointly referred to as the *neural network estimators* throughout the thesis, the considered working hypothesis is defined as follows:

Hypothesis 1 (H1): *The neural network estimators are able to extract information from the return history of the assets. Based on their estimates, they can outperform the baseline estimator in terms of Sharpe ratio by reallocating the portfolio according to the optimized portfolio weights at every trading day without considering trading costs.*

However, it is likely that the neural network estimators result in significantly increased trading volumes, as in contrast to the EWMA estimator, vast differences in the return estimations are possible from one observation to the next, potentially leading to a drastic change in optimal portfolio weights. Thus, the trading volumes are measured in the daily rebalancing case, and it shall be analyzed if common measures used to decrease trading volumes such as a reduced rebalancing frequency or intertemporal smoothing of the optimized portfolio weights (Würtz et al., 2009) can mitigate the potential increase. The following secondary hypothesis, therefore, is also considered:

Hypothesis 2 (H2): *In conjunction with common measures to reduce trading volume, the neural network estimators are able to outperform the baseline estimator in terms of Sharpe ratio in the presence of trading costs.*

3. Methodology

Having covered the theoretical background of the thesis and the hypotheses it seeks to examine, the general structure of the analysis carried out is presented in the following, covering the data set itself, the covariance estimator as well as the neural network and optimization implementations.

3.1. Data

The analysis was performed on data obtained from the data library of Kenneth R. French (Fama & French, 2019) which itself uses stock price data from the *Center for Research in Security Prices (CRSP)* and interest rate data provided by *Ibbotson Associates*. The data set contains synthetic daily returns of a market asset, a risk-free asset, a *high minus low* asset, and a *small minus big* asset. The construction of these assets follows Fama and French (1993). The market asset reflects a value-weighted mean of US stocks listed on the NYSE, AMEX, or NASDAQ. The risk-free asset represents the return on the one-month US Treasury bill. A detailed description of the composition methodology is also available at the previously mentioned data library. In order to keep the scope of the analysis on the comparison of expected return estimators, only the market and risk-free asset are considered, reducing the setting to the two-asset case. In line with the naming conventions of the data set, the market asset is henceforth abbreviated as *Mkt*, while the risk-free asset is referred to as *RF*. The date range covered spans from 1926-07-01 to 2019-06-28, thus covering 93 years with a total of 24,515 observations. In terms of pre-processing, the data set was first analyzed for missing or implausible data points. Since none were observed, no data cleaning was required. The return of the market asset r_{Mkt} is not provided directly, but as excess over the risk-free return. Thus, the return of the risk-free asset must be added to reflect the total returns obtained by an investor in the asset. A first look at the total return indices associated with the asset returns reveals that, over the long observation period, both assets performed as one would expect based on the asset classes. As shown in Figure 5 on a logarithmic ordinate, the market asset shows significant long term growth accompanied by substantial risk and even years with consecutive large drawdowns. Notably, periods of extended upward trends are interrupted by financial crises like the Great Depression from 1929-1939, the oil crisis of 1973, the burst of the dot-com bubble in the early 2000s, the global financial crisis of 2007-2008 and the subsequent European debt crisis. The distribution of these events ensures that the neural network estimators are exposed to different market environments in both the training and the testing phase and therefore, potentially capable of finding subtle patterns in the market returns. On the other hand, the risk-free asset by its nature shows a very smooth return profile. However, the yield of the risk-free asset has undergone significant changes during the time span of the data set. Starting at a relatively low yield of 0.9 basis points per day, a period of prevailing increases in the return on the risk-free asset followed, reaching a peak of 6.1 basis points per day in 1981. Afterward, the yield decreased again, incidentally back to the exact initial yield of 0.9 basis points per day, resulting in a sigmoid-like total return graph. While the EWMA estimator eventually follows both upwards and downwards trends, it remains to be seen if the neural networks could improve on the baseline estimator if the models were to be trained on data stemming from a regime with predominantly increases in the risk-free rate but tested on data with predominantly decreases in the risk-free rate. This could be the case if the regime switch aligns with the split

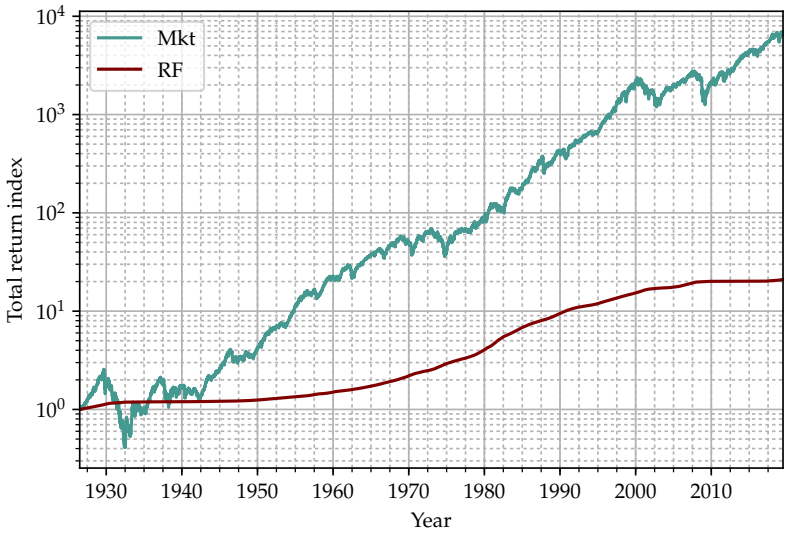


Figure 5: Total return index of assets

of training and testing data.

Next, the asset returns were analyzed for autocorrelation, with the results illustrated in Figures 6 and 7.

As expected, over the full observation period the sample *autocorrelation function (ACF)* is close to zero for the returns of the market assets, with the highest value of 0.067 observed at a lag of one day, as shown in Figure 6. While some of these autocorrelations are significant at the 5% level when compared to the confidence intervals obtained from Bartlett's formula (Bartlett, 1946), they are not stable over time, exhibiting both negative and positive estimates for different subperiods of the return series. This is shown in Figure 7, where the rolling one year (251 day) sample autocorrelation function for a lag of one day on the market asset returns is plotted over time. On the other hand, the risk-free return is by construction almost perfectly positively autocorrelated for all analyzed ten lags. For reference, a correlogram for the risk-free asset is thus provided in the Appendix. While not surprising, this analysis allows for an assessment of the estimators' prediction performance. Since the risk-free asset returns are highly positively autocorrelated, a reasonable prediction is always provided by the previous return. Thus, the differences in the estimates are expected to be small. For the market asset, even though ex-post linear relations to previous returns can easily be identified, finding these linear (or also nonlinear) relationships ex-ante is a different challenge, leading to an expectation of more substantial differences in the estimates.

3.2. Covariance Estimator

While the varied factor in this analysis is the expected return estimator, it is also necessary to estimate the covariance matrix to perform the portfolio optimization. The sample covariance matrix can be a suboptimal estimator when the number of assets approaches the number of observations per asset (Ledoit & Wolf, 2003, 2004). However, in the case at hand, the number of assets is small, and the return history has the same length for all assets. Thus, the sample covariance could be considered a valid estimator. In fact, Lopez and Walter (2002) find that in a Value-at-Risk context, similar to the optimization approach considered in this thesis, simple covariance estimators are not inferior compared to more complex estimators like GARCH models. Especially highlighted by them is the EWMA covariance estimation, which again overweighs more recent observations compared to older ones. Thus, the covariance was estimated using an EWMA model with a smoothing parameter of $\lambda = 0.94$. For daily data, this value was suggested by J.P. Morgan (1996) as well as Walter and Lopez (2000). The resulting annualized standard deviation estimates are shown in Figure 8. The standard deviation of the market asset generally lies between 5% and 20%, with further elevated levels during periods of financial distress and less frequent spikes reaching almost 80%. While the standard deviation of the risk-free asset is not zero, it is minuscule compared to the market asset and thus indistinguishable from zero when viewed on the same scale. The same applies to the covariance of the market asset and the risk-free asset. While the annualized figures are used to provide a more intuitive interpretation, one should note that these values are obtained by scaling the daily volatilities with the factor $\sqrt{251}$. However, this is only accurate when logarithmic returns are used and remains an approximation for the discrete returns used in the portfolio optimization context of this thesis. Therefore, all figures displaying annualized volatilities are replicated using

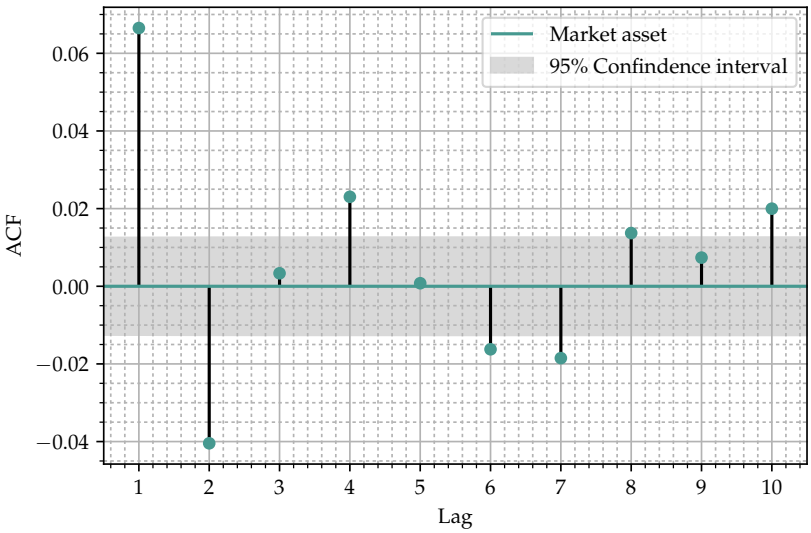


Figure 6: Autocorrelation analysis – Full sample market autocorrelation

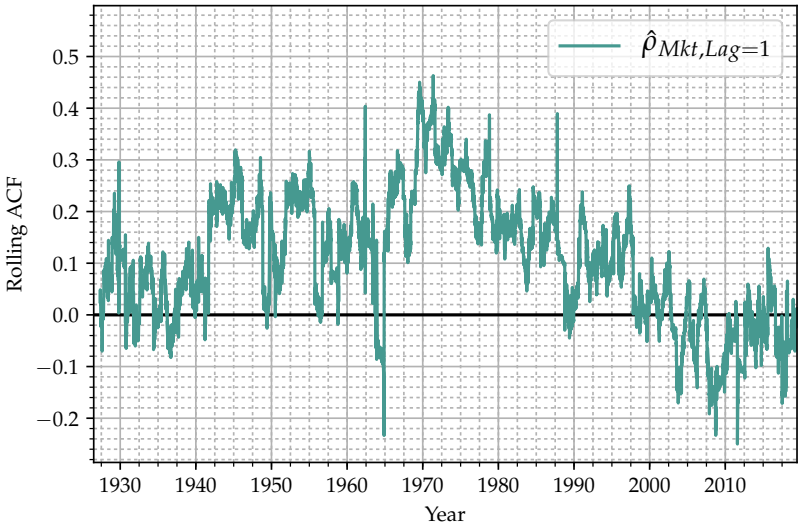


Figure 7: Autocorrelation analysis – Rolling one year market autocorrelation

unscaled daily values in the Appendix.

3.3. Neural Network Implementation

The neural network estimators are implemented in Python using the *TensorFlow* machine learning library (Abadi et al., 2015). More specifically, the high-level application programming interface *Keras* was used within TensorFlow. The following section introduces the three chosen network implementations and describes the considerations behind the choice of the model *hyperparameters*. While model parameters are the weights and biases that are optimized during model training, hyperparameters are any additional configuration factors that themselves influence the training of the neural network, such as the size and number of hidden layers as well as the batch size. Among other approaches, these hyperparameters can be optimized by performing a grid search on a set of reasonable hyperparameter combinations, which can lead to additional improvements in the model performance. It is important to underline, however, that the scope of this thesis is to analyze the general suitability of neural network estimators in the context of dynamic portfolio optimization. Thus, hyperparameter optimization was not performed before the main analysis but is nev-

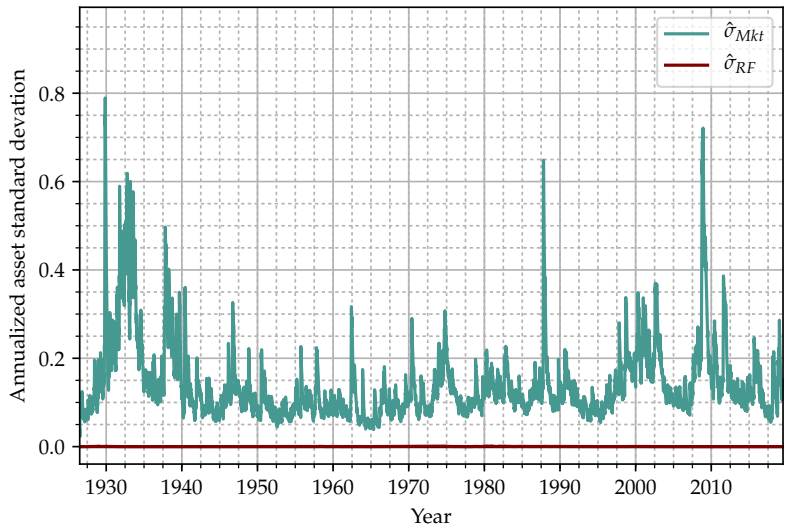


Figure 8: Annualized asset standard deviation over time

ertheless considered in the sensitivity analysis. Rather, sensible hyperparameters were chosen based on related research and best practices.

3.3.1. Multilayer Perceptron

Before implementing an MLP network, the input features have to be selected. As the EWMA estimator only contains information from historical returns, the neural networks also were constrained to this information set. However, while the EWMA estimator encompasses information from all previous observations, a very large number of lagged daily returns would swiftly lead to a dramatic increase in model parameters for the MLP. Thus, a reduced number of features encompassing information about the recent returns had to be chosen. Since the EWMA estimator predicts positive returns if recent returns were also positive and vice versa, a strategy based on this estimator can be considered a *momentum strategy*. Momentum strategies invest in assets which have performed well (compared to other assets or evaluated only on their own return history) over a fixed window of recent observations, usually the last year. These strategies were analyzed by numerous researchers, such as Antonacci (2014); Asness, Frazzini, Israel, and Moskowitz (2014); Baz, Granger, Harvey, Le Roux, and Rattray (2015); Goetzmann and Huang (2018); Hurst, Ooi, and Pedersen (2017); Jegadeesh and Titman (1993); Lempérière, Deremble, Seager, Potters, and Bouchaud (2014); Swinkels (2004), with most finding at least partial support for the strategy. Closely related to the topic of this thesis, Lim et al. (2019) even combined a momentum strategy on futures contracts with neural network estimators and concluded that the neural network estimators outperformed the traditional signals. In order to reflect the informational content of these momentum strategies, the trailing one year return (251 days), one month return (21 days), one week return (5 days), as well as the most recent daily return, are used as features. Additionally, the rolling one month sample standard deviation is added to provide the model with information about the current level of return uncertainty.

Next, the number of hidden layers and their node count had to be decided upon. Given the relatively small size of the data set compared to other machine learning tasks, the model was also kept compact. Goodfellow et al. (2016) find that in different settings, deeper networks provided better generalization. Additionally, Di Persio and Honchar (2016) find that for a similar application in prediction of financial time series, two hidden layers were optimal. Thus, two hidden layers were also used in this model, as shown in the schematic representation in Figure 9.

The number of nodes in each hidden layer was chosen next. A typical default parameterization chooses a number of hidden nodes between one and two times the number of input nodes (Berry & Linoff, 1997; Blum, 1992). Conventionally, the number of nodes are often chosen as powers of two, as this allows for broad coverage of the search space when multiple node counts are compared. Sticking with the convention, with five input nodes, the number of nodes in the hidden layers was set to eight.

For the activation function in the hidden layers, the ReLU function was used. This activation function has numerous theoretical and empirically founded advantages. One example is that its activations are linear and thereby unsaturated for positive inputs, making models using this activation function easier to train (Goodfellow et al., 2016). Additionally, this property also mitigates the van-

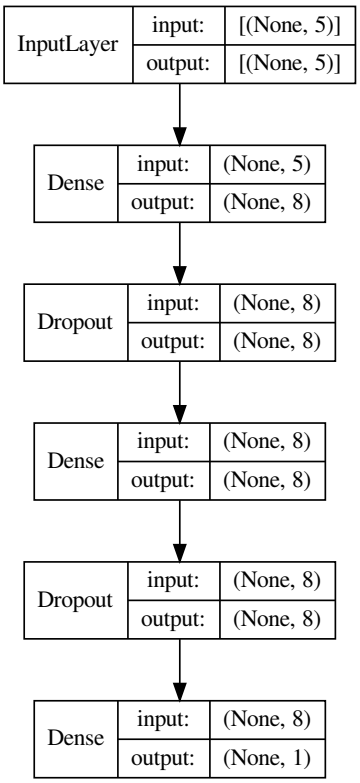


Figure 9: MLP implementation

ishing gradient problem as described by Glorot, Bordes, and Bengio (2011) who also found that ReLU activations more closely reflect the principle of operation of neurons in the brain compared to other popular activation functions. While these considerations led to the selection of the ReLU activation function for the hidden layer, a linear activation was necessary for the output layer, since the prediction of returns is a regression problem.

The number of trainable model parameters in a neural network can often be large compared to other statistical models. Intuitively, it makes sense that neural networks with many parameters can more easily overfit smaller data sets. This was also observed by Srivastava, Hinton, Krizhevsky, Sutskever, and Salakhutdinov (2014), who proposed the implementation of *dropout* layers into the model. A dropout layer randomly deactivates some of its nodes (along with their corresponding weights) with a pre-specified probability during the forward pass and the backpropagation steps of the gradient descent algorithm. The underlying idea is that whenever a set of nodes is deactivated, the other nodes each learn an internal representation of the model, thus reducing a co-adaption of neighboring nodes. This technique was also described by Hinton, Srivastava, Krizhevsky, Sutskever, and Salakhutdinov (2012), suggesting a dropout probability of 0.5. It is therefore often considered best practice to add a dropout layer when implementing a neural network. The chosen model architecture thus introduces dropout layers, as shown in the network architecture schematics in Figure 9. Note that in this representation, the dropout layer does not have its own nodes, but rather sets each node of the preceding dense layer to 0 with probability 0.5. Additionally, it should be noted that the fully connected layers of the MLP model are referred to as *Dense* layers in the Keras framework. Thus, for each sample, there are five input values which are passed through two hidden layers with eight nodes each. Finally, the output layer has a single node since only the return on the next day is estimated. The *None* values refer to the batch size, which is not set yet during the model construction and is further addressed below.

3.3.2. Convolutional Neural Network

For the CNN implementation, schematically depicted in Figure 2, a new set of input features had to be decided upon.

Since the extraction of higher dimensional features is at the core of CNNs, rather than one month, one week and one day return, all 21 lagged daily returns of the preceding month of a given observation were used as features.

Following this argument, the rolling one month standard deviation was also removed from the inputs since this information too could be learned by the model. The one month period was chosen in particular since the half-life and center of mass of the baseline estimator also lie in this region with 17 and 24 days, respectively. However, the rolling one year return was kept as an input in order to retain the longer term information content without dramatically increasing the model parameters through the inclusion of the full year of lagged daily returns. Thus, the total number of input features for each observation is 22. For the convolutional layer, or *Conv1D* as termed in the Keras framework, following Di Persio and Honchar (2016), the size of the filter was chosen to be $[3 \times 1]$ with a stride of $s = 1$. While the number of filters in their analysis was 32 and 64 respectively, considering the size of the data set and the comparability of the number of parameters across the different models of this thesis, only 16 features were used. It should be noted that the Conv1D layer in the network outputs one vector of nodes for each filter, as indicated by the output of an additional dimension by this layer. Furthermore, since no padding is used, the size of the convolutional layer decreases from $[22 \times 1]$ to $[20 \times 1]$ per filter, as per Equation 23. Considering all filters, the total shape of the convolutional layer thus is $[20 \times 16]$. A max pooling layer follows, which uses an input size of $[2 \times 1]$ and stride $s = 2$, thus reducing the number of nodes by 50%. This parameterization again follows Di Persio and Honchar (2016). After the technically required flatten layer, the abstracted features are combined through a fully connected layer with 16 nodes and ReLU activation function. Before the final output layer, a dropout layer is added to counteract overfitting.

3.3.3. Long Short-Term Memory Network

For the LSTM architecture shown in Figure 4, the input layer again had to be created differently from the previous models since it allows the network to create its own internal representation of the importance of previous returns.

At each point in time, the individual daily returns of the previous month (21 days) are given to the model as inputs, again encompassing the half-life and center of mass of the baseline estimator. This window is shifted one day ahead at each new training sample. Since these inputs are not 21 different features but rather 21 lagged observations of a single feature, again a second dimension in the input layer is required. Hence, the tuple $(21, 1)$ is used to specify the input shape.

The 16 nodes in the LSTM layer were again chosen to be approximately equal to the number of input nodes. In this model, the number of nodes describes the length of the vectors used in its gates and the cell state. To allow for further long term dependencies, it is crucial not to shuffle the data set before training, which otherwise is the default implementation in Keras. While other model implementations treat each training sample independently with the input and output values they provide, the very definition of RNNs requires the sequences to be ordered for both training and testing. When it comes to regularization techniques in an RNN network such as the LSTM model, it is possible to not only conceptualize dropouts on the direct path from the input to the output layer, but also on the gates which control the influence of previous hidden layers. Indeed, this was noted by Gal and Ghahramani (2015), who proposed theoretically founded dropouts for RNNs and provided empirical evidence for their usefulness. Thus, in the LSTM model, recurrent dropouts are used in addition to the previously introduced dropouts. Only the regular dropout layer is shown in the schematic, as the recurrent dropouts are included in the LSTM layer in Keras.

Even though the architecture of the LSTM model might look simpler compared to the MLP and CNN models, it is in no way inferior in its capability to represent complex dependencies through its internal gate structure and the transfer of information between samples.

3.3.4. Model training

In order to train the models, the data set first needs to be split into train and test data sets. It is common practice to use the older part of the data as training data and the more recent data as test data in order to improve the likelihood that the performance of the model in an out-of-sample implementation would be closely reflected by the test data performance. The ratio by which to divide the data set is dependent on the application and especially the size of the data set. While on vast data sets the training data can encompass up to 99 percent of the total samples, smaller data sets can require a relatively larger testing set, such that equal splits of the data are also not uncommon (Ng, 2019b). In the data set at hand, a split of 60 percent for the training data and 40 percent for the test data was chosen. Theoretically, this would leave 14,709 observations for the training data set and 9,806 observations for the test data set. However, as shown in Figure 12, not all of this data can be used for training and testing. At the beginning of the training phase, the data required to obtain the first set of input features cannot be used as training examples. If for example, the trailing one year return is one of the input features, this data is not available for data points up to one year into the data set. At the very end of the data set, one also needs to account for the output values. If the prediction target is the return on the next day, then a single observation at the end needs to be omitted from the test data set, as otherwise, no observed return for the prediction on the very last day is available. If returns for longer time horizons are predicted, more observations need to be spared for validating the predictions. Furthermore, it is especially important to separate the training and testing data by a gap with the combined length of the data points required for the input features and the number of days predicted

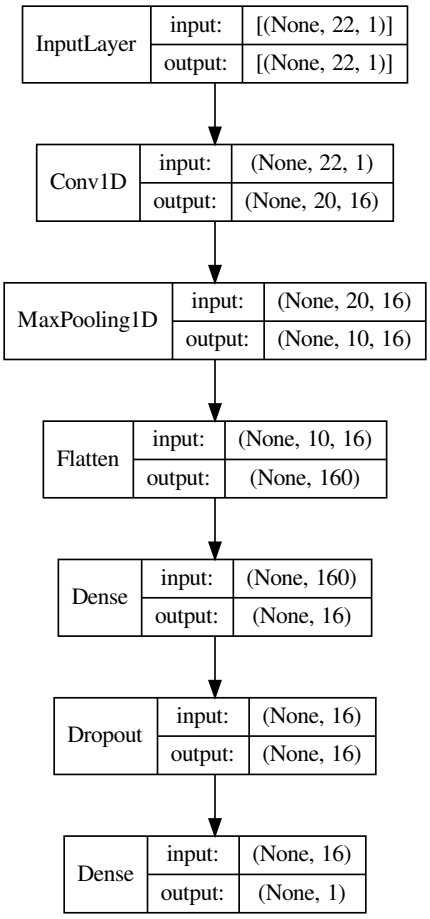


Figure 10: CNN implementation

ahead. The importance of this step stems from the phenomenon that neural networks are outstanding in finding subtle information about the test data within the training data. This is called *information leakage*. In particular, it needs to be assured that the test data does not contain any information the network could have seen during the training phase. Continuing the example of trailing one year returns used as a feature, if no gap between the train and test data sets were present, the returns of the final year of training observations would be part of both the train and test data set and thus contribute to information leakage. A further overview of information leakage is given by Kaufman, Rosset, and Perlich (2011). Removing the observations in question leaves 9,552 samples or approximately 38 years for backtesting the neural networks against the baseline estimator.

While the EWMA estimator is agnostic to the size and variance of the inputs, neural networks can be affected significantly by inputs of varying scales. In particular, if inputs are on different scales, the loss function of the neural network is dominated by the larger features (Theodoridis & Koutroumbas, 2009). Conceptually, for unscaled features, the minima of the error surface become more elliptical, with the gradient descent algorithm potentially oscillating towards the optimum in an inefficient manner. The increased convergence efficiency on scaled features was also noted by Ioffe and Szegedy (2015). Therefore, it is beneficial to scale the input and output data for the training and testing data set to a comparable range. Since the resulting estimates are used in a mean-variance optimization setting, the commonly used method of *z-score normalization* was used to ensure the input features do match in these dimensions. Formally, for each feature x_i with given sample mean $\hat{\mu}$ and sample standard deviation $\hat{\sigma}$, the normalized feature x_i' was computed as

$$x_i' = \frac{x_i - \hat{\mu}}{\hat{\sigma}} \,. \tag{34}$$

However, as suggested by Müller and Guido (2016), scaling the data can also be another source of informa-

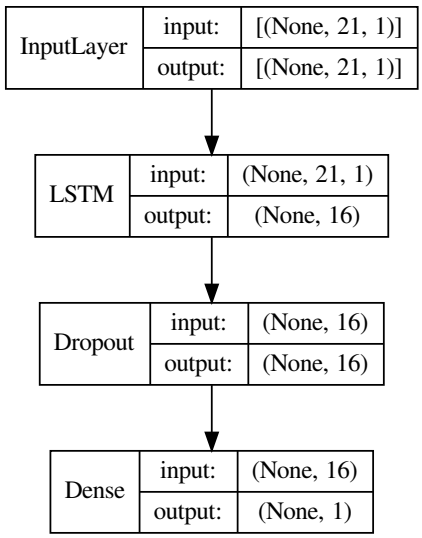


Figure 11: LSTM implementation

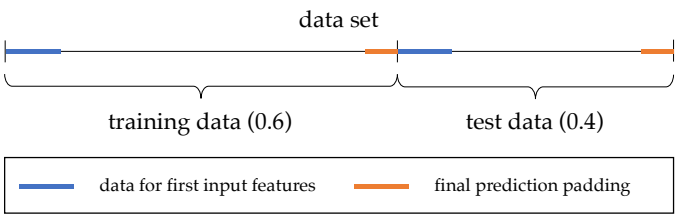


Figure 12: Schematic representation of the train/test split

tion leakage. If the data set is scaled before being split, implicitly the training data contains the mean and standard deviation of the full sample even though this information would not have been available at that time. Thus, the training data needs to be scaled first based only on the estimates obtained on these samples. Next, the test data can be scaled by using the same mean and standard deviation estimates. The same scale is then also used when re-scaling the estimates after training and prediction have been performed.

Finally, before training the model, a batch size had to be chosen. As discussed in section 2.2.1, this parameter can influence the convergence speed and probability of the model considerably. A batch size of 32 is recommended by Bengio (2012) as well as Masters and Luschi (2018). Their research shows that batches in this size region provide a good trade-off between training stability and model generalization across different applications. Thus, this value was chosen as a starting point.

Under consideration of these implementation details, the models could now be trained. The return of the following trading day was used as the output, and the neural networks were trained to minimize the mean squared error with respect to these targets. Computation time was decreased significantly by training the models on a graphics processing unit (GPU) rather than a central processing unit (CPU), with the particular implementation making use of the optimized linear algebra subroutines provided by the Nvidia CUDA framework (Nickolls, Buck, Garland, & Skadron, 2008). Each model was trained for 30 epochs, i.e., the training set was looped over 30 times by the gradient descent algorithm. It was monitored that the loss was decreasing during the training phase (indicating convergence on the training data) and that the loss on the test data was not significantly larger than the loss on the training data (indicating overfitting). It should also be noted that for training the neural networks, a slightly improved version of gradient descent was used, namely the Adaptive Moment Estimation algorithm (*Adam*). This optimizer was suggested by Kingma and Ba (2014) and has since found wide adoption in the field. Compared to the single, fixed learning rate param-

eter α in the introduced GD algorithm, Adam uses varying learning rates for each parameter which are updated based on exponential moving averages of the first and second moment of recent gradients. It is the default optimization algorithm recommended by Ruder (2016).

3.4. Optimization Implementation

When implementing the portfolio optimization, either the approach of maximizing the portfolio return for a given level of risk or the minimization of risk for a given level of return had to be chosen. As seen in Figure 8, the risk of the assets was comparatively stable over time. By contrast, market returns are much harder to predict, and therefore the optimized portfolio could deviate significantly from the targeted return. Additionally, assuming a constant relative risk aversion was found to be a good representation of some investors' utility functions (Chiappori & Paiella, 2011; Wakker, 2008) and thus a constant level of targeted risk for the portfolio is reasonable. With that in mind, the optimization with respect to a prespecified level of risk was selected. In order to reflect different degrees of risk aversion, multiple risk targets were considered. The lowest volatility target was chosen at 2%, well below the lowest volatility estimate for the market asset at over 5% and thus always requiring a portfolio composed of both assets whenever the expected return of the market asset is above the expected return of the risk-free asset. Additional risk targets were placed in increments of 2.5 percentage points, covering a range up to 19.5% annualized volatility, the upper end of the market volatility during most of the observation period.

To obtain the optimized weights necessary for the analysis, the portfolio optimization had to be performed for each date where all estimators delivered estimates (the EWMA estimator does not need training data, but since this split was required for the neural networks, the intersection of the estimation periods had to be used). This included 9,552 dates for which the optimization needed to be performed. For the EWMA estimator, the same smoothing parameter of $\lambda = 0.94$ as for the covariance estimator was chosen. Additionally, alongside the baseline and neural network estimators a *RANDOM* "estimator" was added as a point of reference. This model uses the sample mean and standard deviation of the assets during the training period to parameterize a normal distribution from which a sample was drawn for each testing observation. With the five estimators and the eight considered levels of risk tolerance, this equates to a total of 382,080 optimizations. Conceptually, the optimization seeks to find weights $\mathbf{w}_{target}$, which maximize the expected return for a given level of risk $\bar{\sigma}$, as shown in the optimization problem 2. However, solving this linear program with quadratic constraints was not easily implementable in the chosen programming framework. Thus, the efficient portfolio for the given risk level was obtained by repeatedly minimizing the risk for a given expected return (optimization problem 1a). In the chosen implementation, first, the portfolio weights with minimal standard deviation $\mathbf{w}_{\sigma_{min}}$ were computed. If the risk target was below the standard deviation of this portfolio, σ_{min}, the optimal portfolio weights were set to $\mathbf{w}_{\sigma_{min}}$, which have an expected return of μ_{min}, respectively. Similarly, the point on the efficient frontier with the highest standard deviation, an investment solely in the asset with the highest expected return μ_{max} was obtained. For risk targets larger than the standard deviation of this upper end of the efficient frontier, σ_{max}, again the weights of this portfolio $\mathbf{w}_{\sigma_{max}}$ were used. It should be noted that investing solely in the asset with the highest expected return marks the upper bound of the efficient frontier only when the no short sale constraint mentioned in 1d is added. This was the case throughout the analysis, as otherwise extreme leverage of the portfolio could not be prevented. Since this is held constant between estimators, it is not expected to influence the overall result.

For risk targets between these upper and lower bounds, first define $\Delta_{\mathbf{w}}(\sigma(\mu)) := \sigma(\mu) - \bar{\sigma}$. The original optimization problem of maximizing the expected return for a given level of risk is now equivalent to finding the μ for which the weights with the minimal standard deviation set $\Delta_{\mathbf{w}}(\sigma(\mu)) = 0$. To find this root, *Brent's method* (Brent, 1973) was used, which stands out due to its high reliability and rate of convergence. This is achieved by combining fast converging methods like the *secant method* with the more reliable *bisection method*. As a starting point, either σ_{min} or σ_{max} was used, depending on which had the lower absolute distance to the risk target. Then, for every iterative change in μ, first the optimization problem 1a was solved, yielding the corresponding efficient level or standard deviation $\sigma(\mu)$. The resulting change in $\Delta_{\mathbf{w}}(\sigma(\mu))$ was then used for the next iteration in Brent's method. Dybvig (1984) shows that the efficient frontier in the presence of a no short sale constraint is continuous but not always differentiable at every point, disproving a conjecture of Ross (1977). However, only continuity is required for Brent's method to always converge to $\mathbf{w}_{target}$ (Brent, 1973). Indeed, the convergence was asserted during all optimizations and could always be achieved using the employed implementation.

A disadvantage of this approach, however, is the multiplicatively increased number of optimizations that had to be performed since every initial return maximizing optimization problem was substituted by multiple risk minimizing optimizations. An overview of the full optimization approach is given in Algorithm 2. By using this algorithm, the optimized weights for all data points, estimators, and risk levels were obtained.

4. Results

In the following chapter, the results of the performed analysis are presented and discussed. First, the case of daily portfolio adjustments (also referred to as *rebalancing*) without the consideration of trading costs is ana-

Compute minimum variance portfolio with σ_{min}, μ_{min} and $\mathbf{w}_{\sigma_{min}}$
Compute maximum return portfolio with σ_{max}, μ_{max} and $\mathbf{w}_{\sigma_{max}}$
if $\bar{\sigma} \leq \sigma_{min}$ **then**
  $\mathbf{w}_{target} = \mathbf{w}_{\sigma_{min}}$
else if $\bar{\sigma} \geq \sigma_{max}$ **then**
  $\mathbf{w}_{target} = \mathbf{w}_{\sigma_{max}}$
else
  $\mu = \underset{\mu \in \{\mu_{min}, \mu_{max}\}}{\arg\min} \; |\sigma(\mu) - \bar{\sigma}|$
  $\Delta_{\mathbf{w}} = \sigma(\mu) - \bar{\sigma}$
  while $|\Delta_{\mathbf{w}}| > \epsilon_{tol}$ **do**
    Adjust μ according to Brent's method
    Minimize portfolio volatility for μ, yielding σ_μ and $\mathbf{w}_{\sigma_\mu}$
    $\Delta_{\mathbf{w}} = \sigma(\mu) - \bar{\sigma}$
  end
  $\mathbf{w}_{target} = \mathbf{w}_{\sigma_\mu}$
end

Algorithm 2: Risk targeting optimization

lyzed. Afterward, two approaches to reduce trading volumes are considered: a reduction in trading frequency and the intertemporal smoothing of optimized portfolio weights. In this setting, the portfolio performances are evluated in the presence of trading costs. Finally, a sensitivity analysis is performed to evaluate the robustness of the results.

4.1. Daily Rebalancing

With the daily rebalancing setting chosen as a starting point, the neural networks were trained, and the optimization algorithm was employed for each observation in the test data set. The results of this analysis are presented in the following.

4.1.1. Predictive Accuracy

Before the performances of the dynamically optimized portfolios are analyzed, the underlying return predictions are assessed. Relating to the error function of the networks, a natural first criterion is the mean squared error of the estimates. An overview of this metric for all estimators and both assets is provided in Table 1. Additionally, the related *mean absolute error (MAE)* is stated to add a measure which is less sensitive to outliers.

Unsurprisingly, all estimators perform substantially better on the risk-free asset compared to the market asset. On the risk-free asset, the CNN network has both the lowest MSE and MAE, with the EWMA estimator as a runner-up in both measures. The LSTM and MLP follow, with the largest error made by the RANDOM estimator. For the market asset, the lowest squared and absolute errors are made by the MLP model, with the LSTM, CNN, and EWMA asset closely behind. The relative difference of the baseline and neural network estimators to the RANDOM model is considerably smaller for the market asset, illustrating the large uncertainty in the returns of this asset. To compare the predictive accuracy of two estimators, Diebold and Mariano (1995) propose a test for non-nested models which can be used across a wide variety of error functions, including MSE and MAE. Diebold (2015) suggests that the underlying assumption of the error differential being covariance stationary often provides a reasonable approximation for economic time series. Moreover, Harvey, Leybourne, and Newbold (1997) propose a bias correction for the test statistic and the Student-t distribution to compare this corrected statistic against in order to improve the properties of the test for smaller samples. Table 2 presents the bias-corrected, two-sided test statistics of predictive accuracy comparisons between the neural networks against the baseline estimator, with the null hypothesis stated as a zero error differential. Again, the RANDOM estimator is added as a reference point. The sign of these test statistics indicates whether the baseline estimator had the lower (negative sign) or higher (positive sign) error, corresponding to the errors stated in Table 1.

Although on the risk-free asset the absolute losses were small, all tests show a highly significant loss differential against the baseline, both for the more accurate CNN model as well as the other, less accurate models. On the market asset, the test statistics generally turn out to be smaller. While the lower error of the MLP is still significant against the baseline for both error measures, the error differentials of the CNN and LSTM model do not show significance even at the 10% level on the MSE metric. For the MAE, the inferiority of the LSTM against the baseline estimator also is not significant, but the more precise estimation of the CNN model is. Again, as expected, the RANDOM model is outperformed on both assets with the highest significance. No model has therefore significantly outperformed the EWMA estimator on both assets and error metrics, with the CNN model coming closest and only one error differential being insignificant. Since the magnitude of the error on the risk-free asset is much smaller, it is likely that the accuracy on the market asset dominates the performance of the dynamically optimized portfolios as statistic significance not always implies likewise economic significance. Additionally, while the considered error measures can give a first assessment of the predictions, their accuracy might not be linearly reflected in the portfolio performances. While a lower prediction error indicates a better fit and performance *ceteris paribus*, two different estimators cannot solely be compared on these measures. This can be demonstrated with a simple example. When considering a day with a positive return in an asset and an even higher estimated return for this asset, the portfolio would still benefit from the positive realized return, yet perhaps in an ex-post suboptimal portfolio composition. If, however, the return was underestimated by the same dis-

Table 1: MSE and MAE of daily return predictions

	MSE		MAE	
	RF	Mkt	RF	Mkt
EWMA	*3.09e-10*	1.18e-04	*1.06e-05*	7.24e-03
MLP	2.69e-09	**1.15e-04**	3.65e-05	**7.17e-03**
CNN	**7.52e-11**	1.17e-04	**6.41e-06**	*7.19e-03*
LSTM	5.66e-10	*1.16e-04*	1.88e-05	7.26e-03
RANDOM	2.80e-08	2.28e-04	1.33e-04	1.15e-02

Bold and italic table entries represent the lowest and second lowest error per column, respectively.

Table 2: Diebold-Mariano test for predictive accuracy against baseline estimator

	MSE		MAE	
	RF	Mkt	RF	Mkt
MLP	-35.182^{***}	3.835^{***}	-71.605^{***}	5.687^{***}
CNN	21.667^{***}	0.764	30.121^{***}	3.049^{***}
LSTM	-19.023^{***}	1.550	-36.320^{***}	-1.125
RANDOM	$\mathbf{-64.953}^{***}$	$\mathbf{-36.036}^{***}$	$\mathbf{-120.705}^{***}$	$\mathbf{-50.786}^{***}$

Bold and italic table entries represent the highest and second highest absolute test statistics per risk target, respectively. Negative test statistics correspond to lower errors by the baseline estimator, positive test statistics correspond to higher errors by the baseline estimator.
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$ for a two-sided test.

tance, perhaps even leading to a negative prediction, the optimization might not have chosen this asset, and the portfolio would forgo the positive return. However, the portfolio performance is not only asynchronously dependent on the prediction error, but the portfolio weights resulting from an optimization are also discontinuous in the estimated returns, as for example two assets with the same estimated volatility are swapped when the asset estimated to have the higher expected return changes.

4.1.2. Risk-Return Characteristics

The portfolio returns based on the predictions are assessed next. Since the portfolios are rebalanced daily, it is assumed that at the end of each day t, the portfolios are adjusted towards the optimized weights based on the predictions obtained from the returns preceding and including day t. Denoting these weights as $\mathbf{w}_t$, the portfolio return on the next day $r_{pf,t+1}$ is obtained by summing the returns of each asset after multiplying them by their respective weights within the portfolio. In matrix notation, the portfolio return is obtained as

$$r_{pf,t+1} = \mathbf{r}^{\mathsf{T}}_{assets,t+1}\mathbf{w}_t \,. \tag{35}$$

Iteratively applying this computation to each day in the test period across both risk targets and estimators yields a total of 40 portfolio return time series. A good overview of the performance characteristics of these time series can be obtained by plotting them in the same risk-return space they were optimized in. This is shown in Figure 13[5], where all portfolio performances, as well as the two underlying assets, are shown in the μ-σ-space. The first thing to note is that even the strategies with the highest risk targets are well below their risk budget, with all but the LSTM model reaching a maximum portfolio volatility of around 10%. Since even the market asset only reached around 17% annualized volatility, it is clear that no portfolio could have achieved the upper limit of the risk targets. Additionally, whenever the expected return of the market falls below the risk-free asset, the portfolio is wholly allocated in the latter asset. These periods further decrease the upper risk limit of the portfolios. Since the underlying assets are essentially uncorrelated, it is natural to observe a risk-return profile above the line connecting the realizations of the market and the risk-free asset. As a notable exception, the RANDOM estimator does not even experience this diversification benefit. However, the risk-return profiles of this estimator by construction showed significant inconsistency between runs, sometimes also exhibiting slightly better or worse realizations. While no model quite achieved the return of the market, the CNN model came closest whilst still provid-

[5]All figures in this chapter showing approximated annualized volatilities are again presented in the Appendix using daily figures.

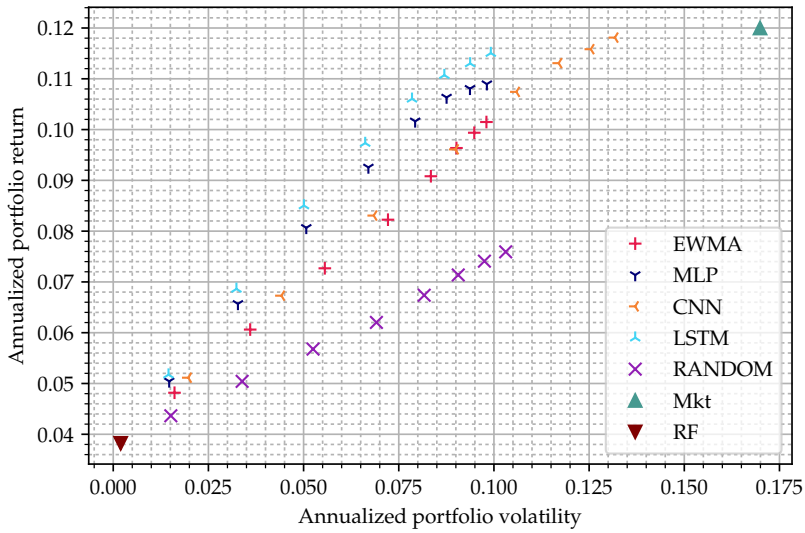


Figure 13: Portfolio performances in μ-σ-space (daily rebalancing, no TC)

ing a lower level of volatility. The LSTM and MLP model further reduce the volatility at their highest risk target, but at an again slightly reduced return. Overall, the MLP and LSTM models achieve higher portfolio returns compared to the EWMA model while staying close to it in the risk space. The realizations using the CNN estimator instead extend the "realized frontier" of the EWMA estimator by providing higher returns in exchange for higher volatility. While the EWMA realizations are inefficient compared to the MLP and LSTM model, the efficiency criterion does not allow for a comparison between the EWMA and the CNN models.

Nevertheless, a closer look at the total return indices of the individual time series is required to assess properties not reflected by their first two moments. The time series for the medium risk target of 12% are shown in Figure 14. Most notably, the MLP and LSTM models achieve substantial outperformance in the beginning, approximately until the turn of the millennium. At that point, the market asset, which is closely tracked by the CNN model, reduces the gap to the MLP and LSTM model, eventually overtaking them for the given risk level. One possible explanation for the decreasing outperformance of the neural networks could be given by the emergence of increased research focusing on the application of these models in the financial industry around that time (see for example Bergerson and Wunsch (1991); Chen, Leung, and Daouk (2003); Deboeck (1994); Gately (1995); Tino, Schittenkopf, and Dorffner (2001)). Practitioners following this research could potentially have lead to the incorporation of this information into the market, reducing the edge these models may have had before. The EWMA estimator also tracks the market asset, however, more loosely than the CNN model. Compared to the market asset, the EWMA model partly avoids the largest drawdowns at the cost of fast recoveries, which is especially apparent at the global financial crisis of 2008, where the market asset dropped sharply by over 50%, followed by an extended period of predominantly positive returns. While the EWMA model fell significantly less during the crisis, it also did not fully experience the following upward trend. In this figure, the improvement of all models over the RANDOM model is also clearly visualized. Looking at the total return indices for the highest and lowest volatility targets generally provided similar insights, and these figures are therefore presented in the Appendix.

The performance of the models during financial downturns is further emphasized in Figure 15, where the risk axis represents the *maximum drawdown (Max DD)* of a portfolio, i.e. the largest loss sustained from a previous high-water mark, rather than its volatility. Again, all models reduce the risk substantially compared to the market asset. What stands out, however, is that the EWMA estimator decreases the maximum drawdown the most, topping at around 26% for the highest volatility target, thereby reducing the risk in this metric by over ten percentage points compared to the neural networks. By construction, the EWMA estimator projects a negative outlook after repeated negative returns, which results in withdrawal from the risky market asset and a shift towards the risk-free asset, thus reducing losses if the market continues to decline. This perhaps is one of the main reasons why the EWMA estimator is often chosen for this application. Conversely, the higher maximum drawdown for the neural networks might be particularly harmful for these estimators. The high degree of complexity of these

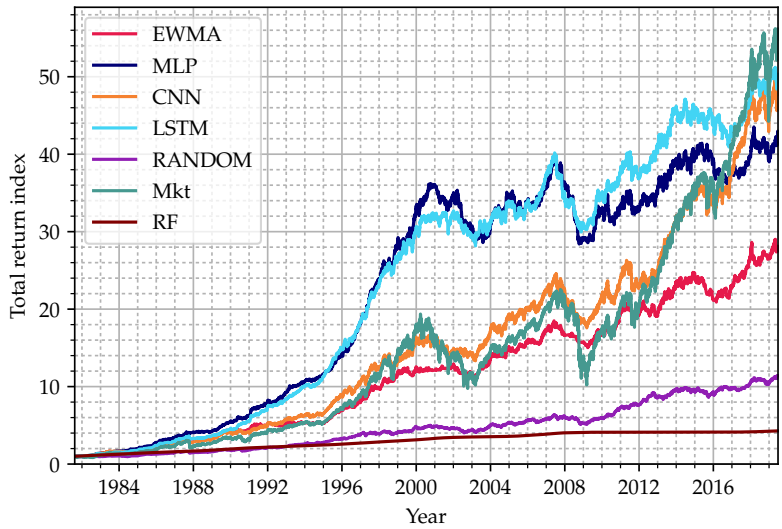


Figure 14: Total return indices for the medium risk target (12% volatility)

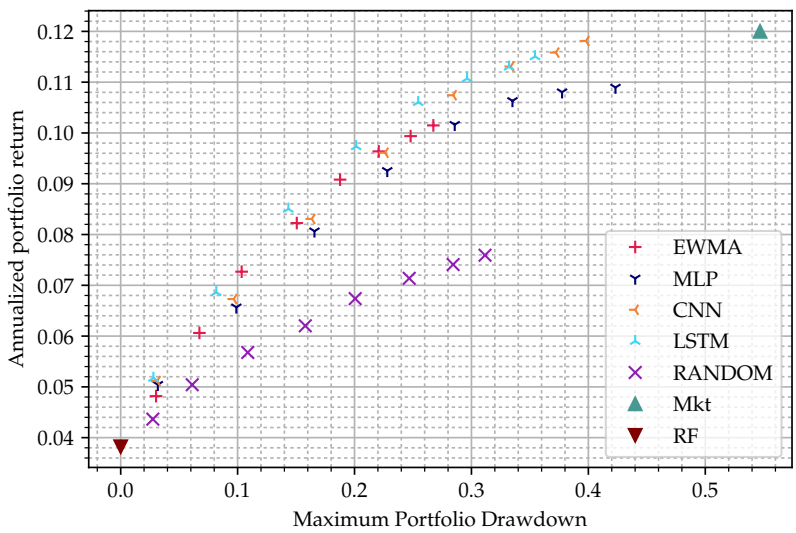


Figure 15: Portfolio performances in μ-Max DD-space (daily rebalancing, no TC)

models compared to more traditional estimators is likely to have an adverse effect on less sophisticated investors from a behavioral point of view, especially since the estimates of these models are not easily traceable. This applies not only to periods marked by large drawdowns but also to periods of relative underperformance against a benchmark.

4.1.3. Portfolio Sharpe Ratios

As the hypotheses are formulated on this basis of the portfolio Sharpe ratios, this metric is considered next. For a more intuitive interpretation, the annualized rather than the daily Sharpe ratios are considered throughout the analysis, with the non-approximated, daily ratios being presented in the Appendix.

Formally, the ex-post Sharpe ratio is defined as $SR = \frac{\bar{r}_e}{\sigma_{r_e}}$, where $\bar{r}_e$ represents the mean return in excess of the risk-free asset and σ_{r_e} represents the volatility of these returns (Sharpe, 1994). Table 3 presents an overview of all 40 Sharpe ratios obtained by the portfolios. As was already apparent from the risk-return figure, the LSTM achieves the highest Sharpe ratio of up to 0.942 on the

lower risk targets. Following the hyperbolic shape of the efficient frontier, the Sharpe ratios decrease for the neural network estimators with increasing risk targets. The Sharpe ratio of the EWMA estimator does not experience this decline and even shows a slight upward trend towards the higher risk targets. Compared to the markets' Sharpe ratio of 0.481 over the same period, all but the RANDOM estimator showed a substantial increase in this metric. While the MLP and LSTM networks achieve higher Sharpe ratios compared to the EWMA estimator across all risk levels, the CNN only leads in the lower risk range.

To assess the significance of these results, a test of Sharpe ratio differentials suggested by Ledoit and Wolf (2008) was conducted. This test is suited for the commonly observed heavy tails of financial time series. It uses a heteroskedasticity and autocorrelation (HAC) robust kernel estimation to compute consistent standard errors on the Sharpe ratio differential. The two-sided test statistics for all estimators against the baseline is given in Table 4, with the null hypothesis stating the equality of Sharpe ratios.

For the MLP and the CNN models, none of the Sharpe ratio differentials are found to be significant. The LSTM model shows significance at the 5% level for the risk targets up to 7% volatility and significance at the 10% level for the risk target of 9.5% volatility, with no significance for higher volatilities. Remarkably, the inferiority of the RANDOM estimator also only shows significance at the 10% level. One should note that even though the RANDOM model has larger test statistics, the LSTM model shows stronger significance for the lowest three risk levels. This is due to the fact that the stated test statistic represents the HAC standard error by which the Sharpe ratio differential is normalized to determine the statistical significance. More specifically, the p-value of the two-sided test is given by $p = 2 * \Phi(-\left|\frac{\widehat{\Delta SR}}{\sigma_{\widehat{\Delta SR}}}\right|)$, where $\Phi(.)$ is the cumulative standard normal distribution function, $\widehat{\Delta SR}$ is the estimated non-annualized Sharpe ratio differential and $\sigma_{\widehat{\Delta SR}}$ is the HAC standard error of this differential. As the LSTM model has the largest Sharpe ratio differential in this lower risk region, it achieves higher significance even though the HAC standard error is slightly smaller.

In light of these results, only a partial support for Hypothesis 1 can be established. While the neural network estimators do show higher Sharpe ratios across most risk targets, few of these performance differentials withstand a test for statistical significance.

4.1.4. Trade Volume

While no trading costs were considered in this case, it is still worthwhile to asses the trade volume associated with these strategies in order to evaluate how drastically the trading volume differs between them and potentially by how much it would need to be reduced to make them viable. Since the portfolios are rebalanced every day, the trade volume is given by the sum of absolute weight differences in the weight vector before and after the rebalancing. At the beginning of day t, the weights $\mathbf{w}_{t,bod}$ are given by the optimized weights of day $t-1$, to which the portfolio was traded. During day t, the weights change through the different returns on the assets, thus at the end of the day, their new weight is determined by their growth, normalized to one by a division through the sum of adjusted portfolio weights for all assets. Formally, for n assets, the end of day weights are given by

$$\mathbf{w}_{t,eod} = \frac{\mathbf{w}_{t,bod} \odot (\mathbf{1} + \mathbf{r}_t)}{\sum_{\nu=1}^{n} w_{\nu t,bod}(1 + r_{\nu t})}, \tag{36}$$

and trade volume at day t is thus being defined as the sum of absolute differences between the end of day weights and the optimal portfolio weights for that day, or formally

$$\mathrm{tv}_t = \sum_{\nu=1}^{n} \left| w_{\nu t,eod} - w_{\nu t,optimized} \right| . \tag{37}$$

One should note that the trade volume defined by Equation 37 is defined in the portfolio weight space, thus making it comparable across time when the portfolio value changes. A switch from a portfolio completely allocated in the risk-free asset to a portfolio completely comprised of the market asset would therefore correspond to a trade volume of 2.0. Table 5 shows the average yearly trade volumes for all estimators and risk targets. Since the EWMA estimator by construction only adjusts its estimates in a smoothed manner, it shows the lowest trade volume across all estimators. Yet still, it too would not be a viable strategy even at low trade costs. While the CNN network shows approximately a doubling, the MLP and especially the LSTM network show a manifold increase in trade volume, with the latter even resulting in higher trade volumes than the RANDOM model. The slight edge in performance of the MLP and LSTM models thus comes in conjunction with an extreme trading volumes which would make them unprofitable even for very low trading costs. These volumes are also indicative of very unstable predictions with little clustering. Thus, in the following, two ways to reduce the trade volume in the presence of these unstable estimates shall be discussed.

4.2. Monthly Rebalancing

The first and perhaps most obvious way to reduce trading costs is to reduce the rebalancing frequency. This represents a trade-off between trading noise in the daily data and forgoing potential profits within this noise. For the neural networks, this changes the way they are being trained since they now do not minimize the error based on the next day's, but rather the next month's (21 days) return. The optimization is then repeated for these models, yielding new optimized weights. Next, the portfolio returns are computed for all risk targets, but the portfolios are only traded towards the optimized weights on the first day of each month.

Table 3: Annualized portfolio Sharpe ratios
(daily rebalancing, no TC)

Vola Target	EWMA	MLP	CNN	LSTM	RANDOM
2.0%	0.624	*0.842*	0.658	**0.942**	0.360
4.5%	0.624	*0.841*	0.659	**0.942**	0.360
7.0%	0.621	*0.838*	0.656	**0.936**	0.353
9.5%	0.611	*0.812*	0.644	**0.894**	0.344
12.0%	0.631	*0.800*	0.654	**0.865**	0.357
14.5%	0.645	*0.778*	0.640	**0.834**	0.366
17.0%	0.645	*0.746*	0.619	**0.798**	0.368
19.5%	0.646	*0.721*	0.606	**0.775**	0.366

Bold and italic table entries represent the highest and second highest Sharpe ratios per risk target, respectively.

Table 4: Ledoit-Wolf test for equality of Sharpe ratios against baseline estimator

Vola Target	MLP	CNN	LSTM	RANDOM
2.0%	0.0088	0.0076	*0.0096***	**0.0100***
4.5%	0.0088	0.0076	*0.0097***	**0.0100***
7.0%	0.0088	0.0076	*0.0097***	**0.0100***
9.5%	0.0088	0.0078	*0.0097**	**0.0101***
12.0%	0.0088	0.0080	*0.0098*	**0.0103***
14.5%	0.0089	0.0084	*0.0100*	**0.0105***
17.0%	0.0089	0.0087	*0.0102*	**0.0107***
19.5%	0.0089	0.0088	*0.0104*	**0.0108***

Bold and italic table entries represent the highest and second highest absolute test statistics per risk target, respectively.
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$ for a two-sided test.

Table 5: Average yearly portfolio trade volume
(daily rebalancing)

Vola Target	EWMA	MLP	CNN	LSTM	RANDOM
2.0%	**6.84**	29.75	*10.10*	42.20	40.42
4.5%	**15.39**	66.96	*22.73*	94.99	90.94
7.0%	**23.62**	103.51	*35.00*	147.04	140.77
9.5%	**30.01**	133.68	*45.26*	191.41	182.82
12.0%	**33.16**	152.85	*52.06*	220.65	210.17
14.5%	34.**82**	163.51	*57.12*	236.82	225.10
17.0%	**36.00**	170.29	*61.08*	246.75	234.08
19.5%	**36.65**	174.67	*63.73*	253.08	239.76

Bold and italic table entries represent the lowest and second lowest trade volume per risk target, respectively.

4.2.1. Predictive Accuracy

A first insight can again be obtained by assessing the predictive accuracy of the monthly return estimates. Table 6 provides an overview of the error metrics of these estimates.

First, it should be noted that these error metrics cannot be compared directly to those of the daily forecasts since the monthly returns are generally different from daily returns in their distributions. Nevertheless, looking at the relative error differences, the results are generally similar to the errors on the daily return forecasts. For the risk-free asset, the baseline estimator provides the best estimates, followed by the CNN model. In contrast, the market asset is again best predicted by the MLP and LSTM models. For this asset, the CNN also performs well, with only the EWMA estimator lagging behind the neural networks. Performing the Diebold-Mariano test against the

Table 6: MSE and MAE of monthly return predictions

	MSE		MAE	
	RF	Mkt	RF	Mkt
EWMA	**2.38e-07**	8.47e-04	**2.97e-04**	2.20e-02
MLP	3.57e-06***	**1.43e-06*****	1.39e-03***	**7.62e-04*****
CNN	*2.54e-07*	8.11e-05***	*3.52e-04****	*3.32e-03****
LSTM	5.97e-07***	*4.22e-05****	6.42e-04***	4.19e-03***
RANDOM	1.17e-05***	1.06e-01***	2.72e-03***	2.54e-01***

Bold and italic table entries represent the lowest and second lowest error per column, respectively.
Asterisks refer to the p-value of the Diebold-Mariano test statistics against the baseline estimator with $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$ for a two-sided test.

baseline estimator also indicates a statistical significance of all error differentials with the exception of the only small edge of the EWMA against the CNN model when comparing the MSE of the risk-free asset. In the table, the significance of the Diebold-Mariano test statistics is also indicated by the asterisks. One should note that these significance indicators refer to an edge of the baseline estimator against the neural networks on the risk-free asset, and vice versa for the market asset. However, as was the case for the daily returns, the overall error on the risk-free asset is much smaller, potentially suggesting a superiority for the neural networks in portfolio performances, which are analyzed next.

4.2.2. Risk-Return Characteristics

A first assessment of the risk-return characteristics of the portfolios can again be obtained by plotting the returns against their volatility, as shown in Figure 16.

Again, the EWMA estimator is able to reduce the portfolio volatility the most, however, only accompanied by lower returns compared to the neural networks. The LSTM network provides higher returns but also higher volatility compared to the baseline estimator, again extending the "realized frontier" of the EWMA model. The MLP and CNN models provide similar risk-return characteristics in this setting, with both models showing both higher risks and returns compared to the EWMA model. These models, unlike the CNN, achieve a better risk-return trade-off compared to the EWMA and LSTM models. Notably, in the monthly rebalancing case, all portfolios are much closer to the line connecting the risk-free and the market asset compared to the daily setting, indicating an outperformance against the market asset as measured by the Sharpe ratio would be unlikely. Keeping in mind that these results were achieved without considering trade costs, the trade volumes of the portfolios are analyzed in the following.

4.2.3. Trade Volume

As can be seen in Table 7, the reduction of trade frequency is extremely effective in reducing the average trade volume of the portfolios. All neural networks show a lower trade volume than the baseline estimator in this setting, potentially also indicating a more stable estimate for monthly returns by these models. Since these trade volumes are in a range where strategies could be profitable when realistic trade costs are considered, the remainder of this section shows results net of trading costs. In fact, trade costs have changed significantly during the observation period and thus assuming them to be constant would not be a good approximation. Jones (2002) assembled average trading costs for stocks in the Dow Jones during the 20th century. Pillared on the presented time series of trading costs, a linear interpolation was performed for the time span from 1980-2000, starting with costs of 90 basis points in the year 1980 and dropping to 20 basis points in the year 2000. Thereafter, the trading costs were conservatively estimated to remain at 20 basis points until 2019. More specifically, these figures represent the relative one-way trading costs consisting of half of the average bid-ask spread plus the exchange commissions. Thus, the return on each day where trading occurs is reduced by the trade volume weighted trading costs. When considering trading costs, the portfolios move vertically in the risk-return space as can be seen comparing Figure 16 to Figure 17. The vertical movement can only be approximated in Table 7 due to the potentially different distribution of trade volumes during the observation period.

Due to its small trading volumes, the MLP achieves the highest returns when trading costs are considered, with only a narrow gap to the CNN model. However, the return distance to the LSTM and EWMA models widens significantly, which due to their similar trade volume continue to be close to each other. For the first time, the random model even underperforms the risk-free asset, illustrating the detrimental effect trading costs can have to unsubstantiated trading strategies. When instead of the

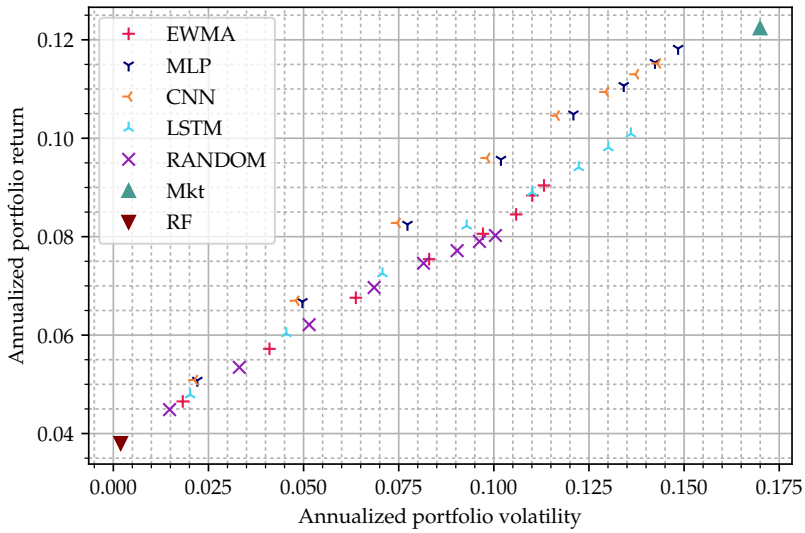


Figure 16: Portfolio performances in μ-σ-space (monthly rebalancing, no TC)

Table 7: Average yearly portfolio trade volume (monthly rebalancing)

Vola Target	EWMA	MLP	CNN	LSTM	RANDOM
2.0%	1.47	**0.53**	*0.71*	1.18	1.93
4.5%	3.32	**1.20**	*1.61*	2.67	4.34
7.0%	5.10	**1.80**	*2.43*	4.08	6.71
9.5%	6.41	**1.93**	*2.78*	5.06	8.76
12.0%	7.07	**1.58**	*2.68*	5.42	10.03
14.5%	7.29	**1.12**	*2.41*	5.42	10.64
17.0%	7.43	**0.85**	*2.30*	5.41	10.96
19.5%	7.50	**0.63**	*2.21*	5.35	11.17

Bold and italic table entries represent the lowest and second lowest trade volume per risk target, respectively.

portfolio volatility, the maximum drawdown is selected as a risk measure, very similar interpretations ensue, with the most notable difference being a slightly wider gap between the MLP and CNN models. For reference, this chart is presented in the Appendix.

4.2.4. Portfolio Sharpe Ratios

Finally, the Sharpe ratios of the portfolios are considered, with an overview given in Table 8.

Unsurprising when recalling the μ-σ plot, the MLP and CNN models achieve the highest Sharpe ratios of up to 0.529 for the MLP and 0.472 for the CNN. The Sharpe ratios of the EWMA and LSTM model, on the other hand, only reach 0.212 and 0.251, respectively. Considering the LSTM model had the highest Sharpe ratios and trade volumes in the daily rebalance setting, it is evident that this model sustained the largest reduction in portfolio returns. Potentially, this model gained its high returns through many small adjustments to the portfolio, which was severely restricted by the switch to a monthly rebalancing schedule. In contrast to the daily rebalancing case, the execution of the Ledoit-Wolf test suggests that the Sharpe ratio differentials of the MLP and CNN models against the baseline estimator are mostly significant at the 1% level, as indicated by the asterisks in the table. Overall, the reduction of trading frequency works well for the MLP and CNN models, providing support for Hypothesis 2, as these models outperform the baseline estimator even in the presence of trading costs. However, the LSTM model does not support this hypothesis under these circumstances but also does not perform worse than the baseline estimator.

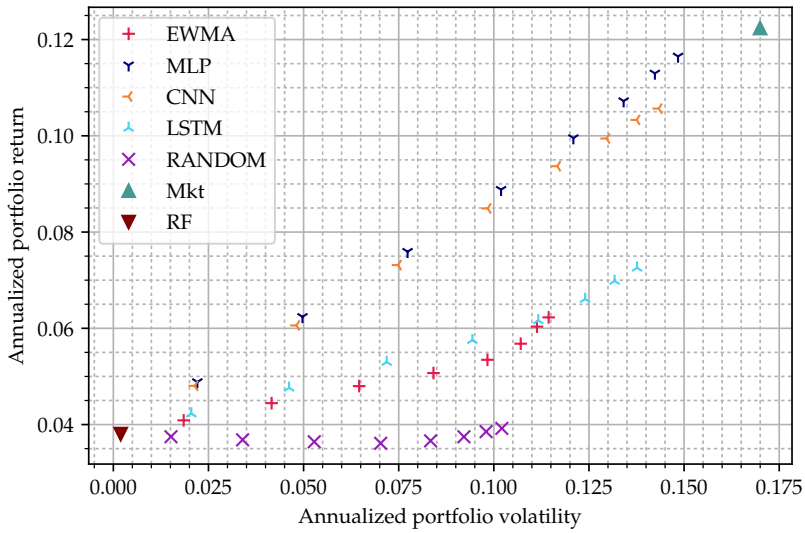


Figure 17: Portfolio performances in μ-σ-space (monthly rebalancing, with TC)

Table 8: Annualized portfolio Sharpe ratios (monthly rebalancing, with TC)

Vola Target	EWMA	MLP	CNN	LSTM	RANDOM
2.0%	0.155	**0.492*****	*0.471****	0.208	−0.036
4.5%	0.155	**0.491*****	*0.469****	0.208	−0.035
7.0%	0.154	**0.490*****	*0.469****	0.208	−0.030
9.5%	0.151	**0.498*****	*0.476****	0.207	−0.027
12.0%	0.157	**0.509*****	*0.477****	0.211	−0.017
14.5%	0.175	**0.516*****	*0.474****	0.226	−0.006
17.0%	0.201	**0.527*****	*0.475***	0.241	0.005
19.5%	0.212	**0.529*****	*0.472***	0.251	0.012

Bold and italic table entries represent the highest and second highest Sharpe ratios per risk target, respectively.
Asterisks refer to the p-value of the Ledoit-Wolf test statistics against the baseline estimator with $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$ for a two-sided test.

4.3. Weight smoothing

As previously established, the low signal-to-noise ratio of historic returns results in a low certainty and high error rate in the forecasts. Similar to the underlying principle of exponentially smoothing previous returns, one could also conceptualize the smoothing of the weights resulting from the portfolio optimizations, such that the weights only change gradually for repeatedly deviating weights. Ideally, this would remove many of the noisy trades while retaining the beneficial shifts in the portfolio weights. One could either apply this smoothing to the weights optimized according to the daily or monthly forecasts or to the forecasts themselves, with the former approach being employed in the analysis. For this, similar to the EWMA estimator, the weights were smoothed in an exponential manner with a smoothing parameter of $\lambda = 0.94$. When smoothing over the optimized weights, the average lag of the weights in the resulting composition clearly is greater than one, and thus, the half-life and the center of mass of the chosen smoothing parameter were considered to find a return estimate better reflecting this property. These lie at 17 and 24 days, respectively, and thus the monthly predictions provide a good fit and the portfolio weights optimized using these forecasts were used. As no fixed trading costs are considered, the portfolios are traded towards these smoothed weights on a daily basis.

4.3.1. Trade volume

In order to assess the effectiveness of this methodology, the trade volumes are examined and presented in

Table 9.

Compared to the monthly rebalancing setting in Table 7, the smoothing of the weights results in a very similar and even slightly higher reduction in trade volumes. In particular, the rank of the models is unchanged, with the MLP model still achieving the lowest trade volume, followed by the CNN, LSTM and eventually the EWMA and RANDOM models. In light of these promising decreases in trade volumes and therefore corresponding trade costs, the analysis can now be focused on the risk-adjusted returns of the portfolios.

4.3.2. Risk-Return Characteristics

Again, the portfolios are plotted in the μ-σ-space for the first assessment of performance. Figure 18 provides this overview.

Like for the trade volumes, the ranking is again unchanged compared to the monthly rebalancing setting. Especially the MLP and CNN models are almost indistinguishable when comparing the two approaches. The LSTM and EWMA, on the other hand, perform better both in terms of risk as well as return when the smoothing method is chosen. The LSTM model can even widen its return margin against the EWMA model, indicating that more of the returns from the unsmoothed daily rebalancing case can be retained compared to monthly rebalancing, whilst showing less overall trade volume. Choosing the maximum drawdown as a risk metric again shows very similar results, implying an approximate proportionality between these measures in both the smoothed and monthly rebalancing settings. The corresponding figure can also be found in the Appendix.

4.3.3. Portfolio Sharpe Ratios

To quantify these implications of the further reductions in trade volumes, the Sharpe ratios of the portfolios in a setting with trade costs are considered next, as presented in Table 10.

Given the similarity in the risk-return figures, it is unsurprising that the Sharpe ratios between the monthly rebalancing and the smoothed rebalancing show a comparable dependency. The MLP model again leads in this metric, closely followed by the CNN model. According to the Ledoit-Wolf test, the outperformance against the EWMA model is slightly less significant in this setting due to the slightly increased Sharpe ratio of the baseline estimator. Up to an annualized volatility target of 12%, the Sharpe ratio differentials are still significant at the 1% level. The higher risk categories remain significant at the 5% level, with solely the 19.5% volatility target of the CNN model showing only significance at the 10% level. The increased Sharpe ratios of the LSTM model, on the other hand, are not enough to constitute significant outperformance against the EWMA estimator as measured by the Ledoit-Wolf test. Conversely, the increased Sharpe ratios of the baseline estimator itself now exhibit an outperformance against the RANDOM model, which is statistically significant at the 1% level. The weight smoothing approach thus underlines the findings of the monthly rebalancing setting and provide further support for Hypothesis 2 by showing significant outperformance of the MLP and CNN model against the baseline. Furthermore, the LSTM model now not only matches but also outperforms the EWMA model, however without the corresponding degree of statistical significance.

4.4. Sensitivity analysis

While the previous sections have given insights into the performance of the employed neural network estimators, it is important to check these results for robustness. Therefore, this section analyzes the sensitivity of these results towards hyperparameter variation, universe expansion, recalibration, and rerun variations.

4.4.1. Hyperparameter variation

Neural networks generally consist of a multitude of model parameters and hyperparameters. As described in Chapter 3, many theoretical and empirical factors were considered when choosing these hyperparameters. This, however, does not guarantee that these hyperparameters are ideal for the specific data set. A variation in these hyperparameters can have a significant impact on model performance, which can be utilized by specifying parts of the data as a validation data set and fine-tuning the hyperparameters based on the results obtained on this data. Alternatively, one can change these hyperparameters after the analysis to assess whether the choices made are stable across a wide range of model configurations or if the results are particularly (un-)favorable only for the narrow set of chosen hyperparameters. Since each hyperparameter can be adapted individually, all possible combinations of these adaptations could theoretically be evaluated. However, each new combination requires a retraining of the estimators and renewed optimization. Therefore, this approach is computationally very costly.

As a narrower approach, for each of the neural networks, both a smaller as well as a larger configuration in terms of layers and nodes was tested. For the larger configurations also a higher dropout probability of 0.6 was used, whereas the smaller configurations only used a dropout probability of 0.4. Finally, the batch size was increased (decreased) from 32 to 64 (16) in the larger (smaller) configuration. The final configurations are also schematically represented in the Appendix. The results of these repeated analyses are depicted in Figure 19. For each of the neural networks, the smaller and lighter markers aptly represent the performances of the smaller models, with the larger and darker markers indicating the larger models. Notably, for the LSTM and CNN models the scaling of the models seemingly did not affect the performance substantially. On the other hand, the MLP model also showed similar performance for the restricted model, but a worse performance for the larger

Table 9: Average yearly portfolio trade volume
(smoothed daily rebalancing)

Vola Target	EWMA	MLP	CNN	LSTM	RANDOM
2.0%	1.01	**0.55**	*0.62*	0.82	1.63
4.5%	2.25	**1.08**	*1.26*	1.76	3.67
7.0%	3.45	**1.44**	*1.77*	2.62	5.68
9.5%	4.42	**1.45**	*1.96*	3.24	7.39
12.0%	5.02	**1.18**	*1.93*	3.54	8.53
14.5%	5.30	**0.86**	*1.82*	3.62	9.15
17.0%	5.46	**0.65**	*1.79*	3.65	9.53
19.5%	5.56	**0.48**	*1.77*	3.65	9.75

Bold and italic table entries represent the lowest and second lowest trade volume per risk target, respectively.

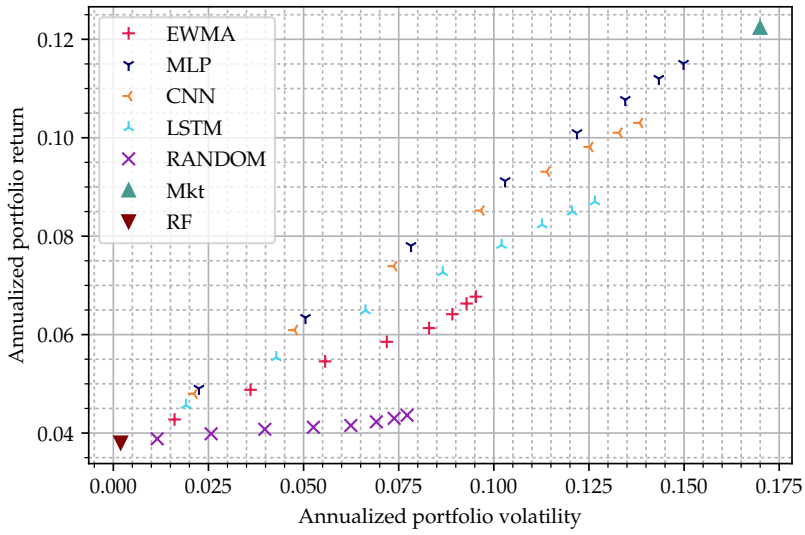


Figure 18: Portfolio performances in μ-σ-space
(smoothed daily rebalancing, with TC)

Table 10: Annualized portfolio Sharpe ratios
(smoothed daily rebalancing, with TC)

Vola Target	EWMA	MLP	CNN	LSTM	RANDOM
2.0%	0.296	**0.491*****	*0.472****	0.402	0.071***
4.5%	0.299	**0.503*****	*0.481****	0.405	0.071***
7.0%	0.297	**0.512*****	*0.487****	0.405	0.069***
9.5%	0.285	**0.517*****	*0.488****	0.400	0.060***
12.0%	0.281	**0.517*****	*0.484****	0.394	0.056***
14.5%	0.293	**0.518****	*0.480***	0.393	0.062***
17.0%	0.305	**0.516****	*0.474***	0.390	0.067***
19.5%	0.312	**0.514****	*0.470**	0.387	0.072***

Bold and italic table entries represent the highest and second highest Sharpe ratios per risk target, respectively.
Asterisks refer to the p-value of the Ledoit-Wolf test statistics against the baseline estimator with $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$ for a two-sided test.

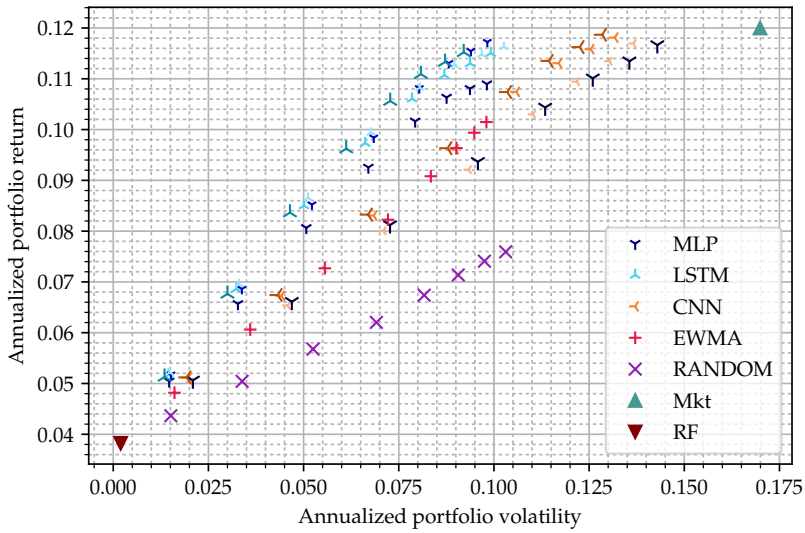


Figure 19: Portfolio performances in μ-σ-space (small and large models, daily rebalancing, no TC)

model. While the larger model did, in fact, show higher returns, these came at the cost of over-proportionally higher volatilities, thus reducing the overall performance from a Sharpe ratio perspective. A potential explanation for the different performance characteristics for this model could be overfitting, as the nodes per layer were doubled, thereby quadratically increasing the parameters per fully connected layer, as apparent when looking at the construction of the weight matrix (compare the sample weight matrix in Equation 5). The addition of another fully connected layer could further have contributed to the overfitting of the model. Overall, it can be concluded that for the data set at hand, the LSTM and CNN models seem robust against changes in the model architecture, with the MLP model showing inferior performance for larger models. Thus, when implementing an MLP model, special attention should be paid to the choice of hyperparameters, and the usage of parts of the data set to validate hyperparameter choices seems advantageous.

4.4.2. Universe expansion

While the presented analysis considered a two-asset case, the implementation can easily be expanded to a multi-asset case. Thus, the two previously excluded assets representing the *high minus low (HML)* and *small minus big (SMB)* factors are added to the estimation and subsequent optimization. This provides further valuable insights, as unlike the risk-free asset, these factor assets are not completely uncorrelated to the market and therefore, can provide a better proxy for a realistic application of portfolio optimization. Figure 20 shows the performances of the portfolios optimized on the extended universe in the μ-σ-space. It is apparent that both the HML and SMB factor assets did underperform even the risk-free asset during the testing period. For approximately half of the volatility of the market asset, the HML return is similar to the risk-free return, and the SMB effectively has no return at all. These full period measures do not necessarily hold true for every sub-period, and outperformance can thus be achieved by shifting the portfolio weights towards assets with good risk-return trade-offs at any given day. In fact, all neural networks estimators seem to be better in this anticipatory shifting of portfolio weights, as the outperformance against the baseline estimator increases substantially. Furthermore, all neural networks achieve higher absolute returns compared to the market asset at the upper end of the risk targets, which was not achieved in the two-asset case. The portfolios' volatilities, on the other hand, are not affected substantially, resulting in greatly improved Sharpe ratios for these portfolios, as indicated by the slope of the "realized frontiers".

Overall, the expansion of the asset universe does not undermine the results of the main analysis based on the two-asset case, but rather provide further support for the usefulness of the neural network estimators. Additionally, the remarkable performance improvements obtained through the addition of two assets which on their own underperformed during the training period opens up further research questions like the inclusion of short assets (i.e., assets which track the inverse performance of a benchmark or index), which are also expected to underperform in the long run in this setting due to the positive expected return of the market, but can provide excess returns when held during times of financial distress. Furthermore, these results justify the expan-

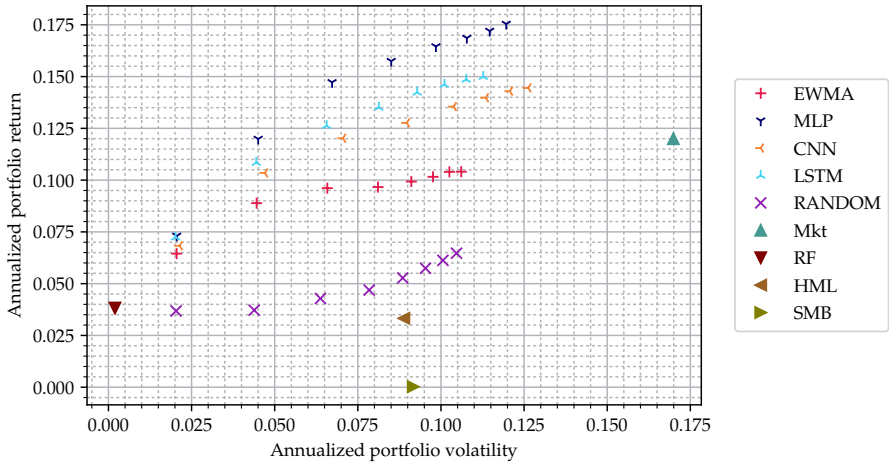


Figure 20: Portfolio performances in μ-σ-space
(all assets, daily rebalancing, no TC)

sion of the analysis across a broader set of asset classes, thereby providing the optimization access to different risk-return profiles, which are more or less advantageous at any given moment. To achieve optimal performance, the portfolio composition, therefore, needs to be adapted over time, a task at which the incorporation of neural network estimators seemingly can be beneficial.

4.4.3. Recalibration

Since capital markets are constantly evolving, even a well-performing model trained on any given data set would become outdated at some point. Therefore, models need to be updated to reflect the information contained in more recent data. Neural networks can potentially deal with this phenomenon by being retrained on an expanding data set, also referred to as recalibration.

Theoretically, one could update the model parameters incrementally using a single stochastic gradient descent step whenever a single new observation is available. This is referred to as *incremental* or *online learning*. It is in so far advantageous as only the gradient for the most recent sample has to be computed, thus making it computationally more efficient. A more comprehensive overview of this technique is given by Gepperth and Hammer (2016).

On the other hand, completely retraining the model from a random initialization using the extended data set is referred to as *offline learning*. This method has the advantage that it is often easier to implement and monitor. However, due to the random initialization and the non-convex error function, it is also possible to reach a different local minimum during the next optimization, potentially causing a more drastic change in model behavior after each retraining. It is also possible to recalibrate the models at less frequent intervals, for example, once per month or once per year.

To assess the possible improvements that can be gained from this technique, a single recalibration step is added to the analysis. After the first half of the test data set is predicted, these observations are added to the training data set. Next, the model is retrained from a random initialization on the extended data, again preventing information leakage by introducing a gap between the (extended) training data and the (remaining) test data. Finally, a prediction on the second half of the test set is performed. The performance of the recalibrated and non-recalibrated models are plotted in Figure 21. Since the recalibration is only relevant for the neural networks, the EWMA and RANDOM estimators are omitted from the figure. The added larger and darker markers refer to the recalibrated models. The overall effect of recalibrating the model is not consistent, with the LSTM model showing only very minor differences, the MLP model showing a substantial improvement in returns accompanied by an under-proportional increase in volatility and the CNN even showing a small reduction in returns at similar volatilitie. Possibly, the different patterns recognized by the models in the additional data points were not equally recurring in the following testing period. Moreover, one should consider whether an extension of the training window is necessarily the ideal approach, or whether a shifting of the training window would be more efficient as this would remove the oldest data which potentially is least relevant for the prediction.

An alternative approach would be to extend the window, but weight each sample according to its recency, thereby overweighing newer observations compared to older ones. An overview of this approach is given by Morantz, Whalen, and Zhang (2008). While different weighting schemes can be conceptualized, following the argumentation of the EWMA estimator itself, an exponential decay of the importance of previous samples could be used as a starting point. In conclusion, the recalibration of the model can substantially affect the performance of the model in the long run but is not a

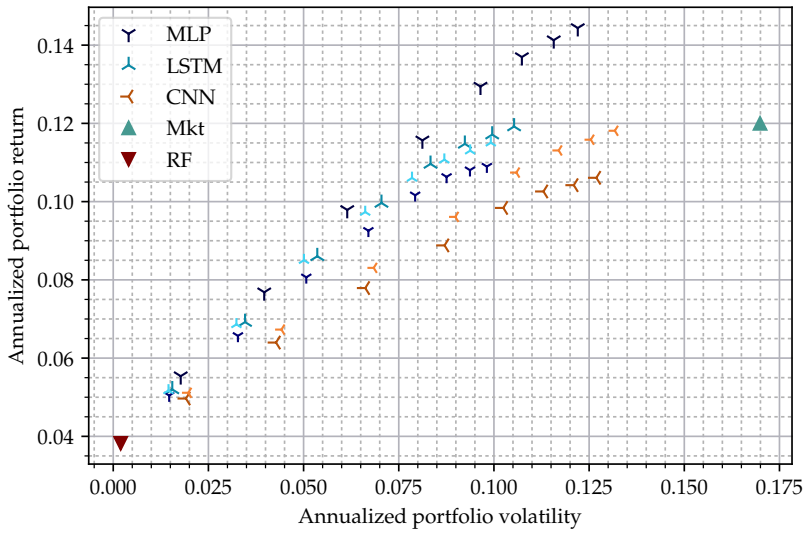


Figure 21: Portfolio performances in μ-σ-space (recalibrated, daily rebalancing, no TC)

guaranteed way to improve performance. The different introduced approaches need to be carefully evaluated when implementing a neural network, especially when considering the computational burden of frequently recalibrating large models.

4.4.4. Prediction consistency

Since the gradient descent algorithm starts with random initialization of the model, it is possible that different local minima are obtained as the objective function is often non-convex. While Du, Lee, Li, Wang, and Zhai (2018) show that overparameterized networks can find the global minimum of the error function, Choromanska, Henaff, Mathieu, Arous, and LeCun (2014) and Goodfellow et al. (2016) conclude that it is generally more likely to obtain local minima close to the global minimum. Furthermore, at the point at which the training process is stopped, as defined by the number of epochs, different model parameters can follow from different paths towards the minimum as a result of the random initialization. Repeating the analysis thus gives insights into how different the obtained optimization results are. Additionally, repeating the analysis supports the reproducibility of results. While commonly a fixed random seed is used in the analysis, this does not guarantee that the outcomes based on this seed are representative for any starting values. Furthermore, it is not easily possible to fix all occurrences of randomness in the interplay of different programming libraries used. Thus, the main analysis with daily rebalancing was repeated three times. Figure 22 gives a first overview of the results, with again the EWMA and RANDOM estimators being omitted. Compared to the initial analysis, the two additional repetitions are again presented with smaller and lighter as well as larger and darker markers. The LSTM model shows the smallest rerun variation across the samples, with both repeated "realized frontiers" closely tracking the initial performances. For the MLP model, one repetition shows higher returns for similar volatilities, whereas the other shows higher returns at higher volatilities, with the overall results being similar. For the CNN model, on the other hand, both repetitions show up to two percentage points lower returns at similar volatilities, marking the largest deviation across the models. Most importantly, the overall impression of model performances remains valid when comparing multiple runs of model training, especially when considering the sensitivity of portfolio optimizations to the estimation of the expected asset returns. This is especially pronounced when the case of daily rebalancing based on single day return predictions is considered where forecasts are particularly unstable. Thus, the analysis of rerun consistency was also performed for the daily smoothed portfolios based on monthly predicted returns. As apparent from Figure 23, the smoothing in combination with the substantially more stable monthly returns, results in almost unaltered portfolio performances across multiple runs.

Overall, the repetition of the analysis provides support for the main findings by corroborating that the observed portfolio performances based on the neural network estimates were not coincidentally obtained from an exceptional manifestation of otherwise clearly worse-performing models. However, for practical applications, variation between runs would obviously pose a greater problem. In this case, one could consider choosing either the return forecasts or the optimized weights based

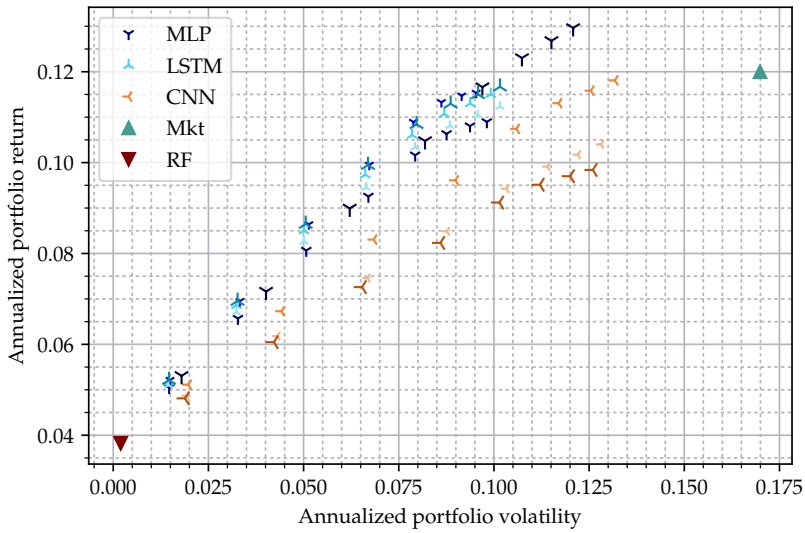


Figure 22: Portfolio performances in μ-σ-space
(rerun consistency, daily rebalancing, no TC)

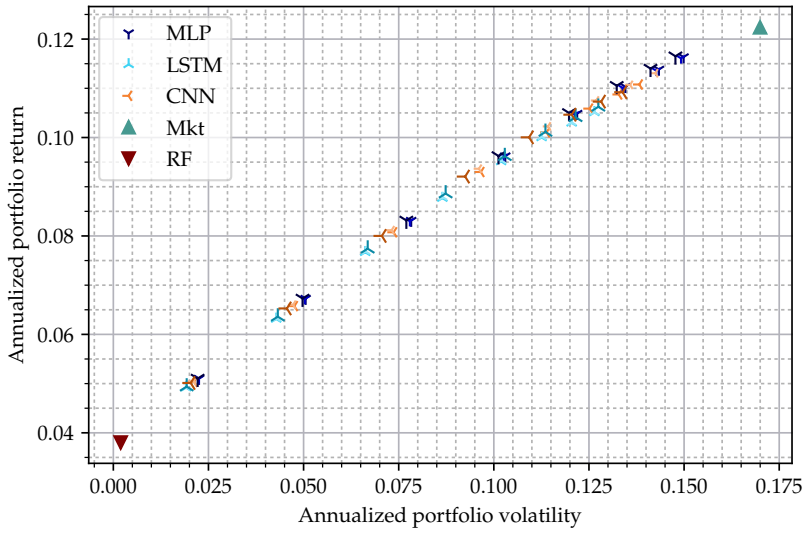


Figure 23: Portfolio performances in μ-σ-space
(rerun consistency, smoothed daily rebalancing, no TC)

on the mean of multiple runs, which would, however, again sharply increase the computational demands of the implementation.

4.5. Discussion

As could be shown in the previous sections of this thesis and the resulting partial acceptance of the formulated hypotheses, neural networks were found to be beneficial in a portfolio selection framework for the data set under examination. However, several challenges arise with the generalization of these results.

Firstly, certain common assumptions in academia employed in this thesis do not entirely reflect the real-world dynamics of financial markets. One of the most common deviations is that portfolios can be optimized after observing the closing prices in real-time, and the resulting allocations are also obtained based on these very prices. Arguably, this is rather a result of data availability and more specifically, the lack of intraday prices, whose presence would allow modeling a time gap between op-

timization and trading prices. However, this most likely does not impact the fundamental interpretability of results, as price differences between optimization and execution prices are not one-sided and likely provide netting effects over time. A more severe impact likely arises due to the negligence of tax effects, as potential savings from tax deferrals are not realized when sell trades on assets with taxable gains are executed. Yet, the baseline model also exhibits this negligence and in fact, boasts even higher trading volumes in the monthly and smoothed rebalancing settings compared to the neural network estimators, thus potentially even strengthening the neural networks' edge. In order to avoid such trades, a buy and hold strategy could also be implemented and indeed would be a viable option to compare against the neural networks in future analyses, especially since the market asset provided the highest absolute return in many settings.

Secondly, implementations of neural networks, especially in the analysis of time series, often are subject to explicit and implicit research biases. Hence, special care was taken to avoid common pitfalls within the process. This includes the prevention of information leakage through complete separation of training and testing data for both feature engineering and data scaling. At the same time, the model hyperparameters were chosen based on academic consensus and best practices rather than repeated iterations on the same data set, which would immediately invalidate the generalization power.

Thirdly, while it lies beyond the scope of this analysis, the additional model complexity brought on by the neural networks can potentially lead to behavioral consequences when implemented in a real-time environment. One example could be a reduced model confidence in times of underperformance, possibly leading to adverse decision-making. This is especially true for neural networks, as their mapping process of inputs and outputs is poorly understood and can be counterintuitive at times (e.g. Szegedy et al., 2013).

Finally, a higher degree of reliability on the results can only be obtained by repeating the analysis for additional data sets reflecting other asset classes, markets and time periods, thereby assessing whether the properties of the neural networks in a portfolio optimization framework retain their validity across the multitude of possible applications found in financial data.

5. Conclusion

After a brief introduction of the dynamic portfolio optimization framework, three types of neural networks with potentially advantageous properties for the estimation of expected asset returns are presented: the Multilayer Perceptron, Convolutional Neural Network, and Long Short-Term Memory Network. The usefulness of these neural networks is then evaluated by using their asset return estimates within a dynamic portfolio optimization framework and an asset universe consisting of a risk-free and a market asset. This thesis finds evidence that in the analyzed data set the neural networks were able to outperform a more traditional EWMA estimator in most settings. In a daily rebalancing scenario and utilizing daily return estimates, the outperformance was mostly insignificant, and the neural network estimates would have resulted in infeasible trade volumes, thus providing only partial support for the corresponding hypothesis. Using monthly return estimates and a likewise monthly rebalancing frequency sharply reduced the trade volumes of the neural networks, resulting in a mostly significant outperformance when historically adjusted trading costs are considered. Applying a smoothing function to the optimized portfolio weights and again using a daily rebalancing frequency results in even lower trade volumes, thus further strengthening the outperformance of the neural networks and yielding an increase in Sharpe ratios of up to 0.2 against the EWMA baseline. Therefore, support is presented for the hypothesis associated with outperformance in the scenarios aiming at reduced trading volumes. The results were further validated through the assessment of model sensitivities towards hyperparameter variations, recalibrations, and repeated executions. Moreover, the analysis of an asset universe extended beyond two assets has shown remarkable results, substantiating the future analysis of other combinations of asset classes, market regions, and time periods.

Besides these naturally emerging research questions, the incorporation of neural network research into the portfolio optimization framework opens a multitude of additional research opportunities. Related to this thesis, weighting schemes for individual observations during the training phase could be considered in order to increase the relevance of more recent samples. Furthermore, instead of modeling each asset's time series individually, returns of other assets could be included as additional features to implicitly reflect the dependency structure between asset returns in the model. Neural networks are not restricted to the use of historic returns either, enabling the inclusion of fundamental or otherwise relevant data into the estimation. Similarly, neural networks can directly be incorporated in the estimation of asset covariances or, taking it one step further, the optimal portfolio weights (and thus performance) could directly be targeted by a neural network. Thereby the moment estimation and optimization steps can be combined in a single architecture, providing a transition to reinforcement learning models. Whenever neural networks are applied in the context of financial data, however, one needs to pay particular attention to the unique challenges arising from these models, such that a compromise of the result validity through repeated analysis on a single data set or information leakage can be prevented. As reinforced by this thesis, financial markets by their very

nature continue to provide a challenging environment for reliable forecasting. Depending on the implementation, advanced models such as neural networks might be able to provide improvements in academic analyses, but one always needs to keep in mind the inherent limitations of forecasts in this setting. For practitioners, the potential adverse effects of additional model complexity should not be underestimated, and if implemented, research into the traceability of the neural network estimates is demanded.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., … Zheng, X. (2015). *TensorFlow: Large-scale machine learning on heterogeneous systems*. Retrieved from http://tensorflow.org/ (Software available from tensorflow.org)

Alpaydin, E. (2010). *Introduction to machine learning*. Cambridge: MIT Press.

Antonacci, G. (2014). *Dual Momentum Investing: An Innovative Approach for Higher Returns with Lower Risk*. New York: McGraw-Hill Education.

Asness, C., Frazzini, A., Israel, R., & Moskowitz, T. (2014, 9). Fact, Fiction, and Momentum Investing. *The Journal of Portfolio Management*, *40*(5), 75–92. Retrieved from http://jpm.pm-research.com/lookup/doi/10.3905/jpm.2014.40.5.075 doi: 10.3905/jpm.2014.40.5.075

Azoff, E. M. (1994). *Neural network time series forecasting of financial markets* (1st ed.). New York: John Wiley & Sons.

Bartlett, M. S. (1946). On the Theoretical Specification and Sampling Properties of Autocorrelated Time-Series. *Supplement to the Journal of the Royal Statistical Society*, *8*(1), 27–41. Retrieved from https://www.jstor.org/stable/2983611?origin=crossref doi: 10.2307/2983611

Baz, J., Granger, N. M., Harvey, C. R., Le Roux, N., & Rattray, S. (2015). Dissecting Investment Strategies in the Cross Section and Time Series. *SSRN Electronic Journal*. Retrieved from http://www.ssrn.com/abstract=2695101 doi: 10.2139/ssrn.2695101

Bengio, Y. (2009). Learning Deep Architectures for AI. *Foundations and Trends in Machine Learning*, *2*(1), 1–127. Retrieved from http://www.nowpublishers.com/article/Details/MAL-006 doi: 10.1561/2200000006

Bengio, Y. (2012). Practical Recommendations for Gradient-Based Training of Deep Architectures. (arXiv:1206.5533v2 [cs.LG]). Retrieved from http://deeplearning.net/software/pylearn2

Bengio, Y., Simard, P., & Frasconi, P. (1994, 3). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, *5*(2), 157–166. Retrieved from https://ieeexplore.ieee.org/document/279181/ doi: 10.1109/72.279181

Bergerson, K., & Wunsch, D. C. (1991). A commodity trading model based on a neural network-expert system hybrid. In *Ijcnn-91-seattle international joint conference on neural networks* (Vol. i, pp. 289–293). IEEE. Retrieved from http://ieeexplore.ieee.org/document/155192/ doi: 10.1109/IJCNN.1991.155192

Berry, M. J. A., & Linoff, G. (1997). *Data Mining Techniques*. New York: John Wiley & Sons.

Best, M. J., & Grauer, R. R. (1991, 4). On the Sensitivity of Mean-Variance-Efficient Portfolios to Changes in Asset Means: Some Analytical and Computational Results. *Review of Financial Studies*, *4*(2), 315–342. Retrieved from https://academic.oup.com/rfs/article-lookup/doi/10.1093/rfs/4.2.315 doi: 10.1093/rfs/4.2.315

Black, F., & Litterman, R. (1992, 9). Global Portfolio Optimization. *Financial Analysts Journal*, *48*(5), 28–43. Retrieved from https://www.tandfonline.com/doi/full/10.2469/faj.v48.n5.28 doi: 10.2469/faj.v48.n5.28

Blum, A. (1992). *Neural Networks in C++*. New York: John Wiley & Sons.

Borovykh, A., Bohte, S., & Oosterlee, C. W. (2017, 3). Conditional Time Series Forecasting with Convolutional Neural Networks. (arXiv:1703.04691v5 [stat.ML]). Retrieved from http://arxiv.org/abs/1703.04691

Brent, R. P. (1973). Algorithms for Minimization Without Derivatives. In *Algorithms for minimization without derivatives* (chap. 3-4). Englewood Cliffs: Prentice-Hall.

Brown, D. B., & Smith, J. E. (2011, 10). Dynamic Portfolio Optimization with Transaction Costs: Heuristics and Dual Bounds. *Management Science*, *57*(10), 1752–1770. Retrieved from http://pubsonline.informs.org/doi/abs/10.1287/mnsc.1110.1377 doi: 10.1287/mnsc.1110.1377

Chen, A.-S., Leung, M. T., & Daouk, H. (2003, 5). Application of neural networks to an emerging financial market: forecasting and trading the Taiwan Stock Index. *Computers & Operations Research*, *30*(6), 901–923. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S0305054802000370 doi: 10.1016/S0305-0548(02)00037-0

Chiappori, P.-A., & Paiella, M. (2011, 12). Relative risk aversion is constant: evidence from panel data. *Journal of the European Economic Association*, *9*(6), 1021–1052. Retrieved from https://academic.oup.com/jeea/article-lookup/doi/10.1111/j.1542-4774.2011.01046.x doi: 10.1111/j.1542-4774.2011.01046.x

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014, 6). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. (arXiv:1406.1078v3 [cs.CL]). Retrieved from http://arxiv.org/abs/1406.1078

Choromanska, A., Henaff, M., Mathieu, M., Arous, G. B., & LeCun, Y. (2014, 11). The Loss Surfaces of Multilayer Networks. (arXiv:1412.0233v3 [cs.LG]). Retrieved from http://arxiv.org/abs/1412.0233

Ciresan, D., Meier, U., & Schmidhuber, J. (2012, 6). Multi-column deep neural networks for image classification. In *2012 ieee conference on computer vision and pattern recognition* (pp. 3642–3649). IEEE. Retrieved from http://ieeexplore.ieee.org/document/6248110/ doi: 10.1109/CVPR.2012.6248110

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011, 3). Natural Language Processing (almost) from Scratch. (arXiv:1103.0398v1 [cs.LG]). Retrieved from http://arxiv.org/abs/1103.0398

Cootner, P. H. (1964). *The random character of stock market prices*. Cambridge: MIT Press.

Cybenko, G. (1989, 12). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, *2*(4), 303–314. Retrieved from http://link.springer.com/10.1007/BF02551274 doi: 10.1007/BF02551274

Deboeck, G. J. (1994). *Trading on the edge: neural, genetic, and fuzzy systems for chaotic financial markets* (Vol. 39). New York: John Wiley & Sons.

Diebold, F. X. (2015, 1). Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold–Mariano Tests. *Journal of Business & Economic Statistics*, *33*(1), 1–9. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/07350015.2014.983236 doi: 10.1080/07350015.2014.983236

Diebold, F. X., & Mariano, R. S. (1995, 7). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253–263. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/07350015.1995.10524599 doi: 10.1080/07350015.1995.10524599

Di Persio, L., & Honchar, O. (2016). Artificial neural networks architectures for stock price prediction: Comparisons and applications. *International Journal of Circuits, Systems and Signal Processing*, *10*, 403–413. Retrieved from http://www.naun.org/main/NAUN/circuitssystemssignal/2016/b482005-303.pdf

Du, S. S., Lee, J. D., Li, H., Wang, L., & Zhai, X. (2018, 11). Gradient Descent Finds Global Minima of Deep Neural Networks. (arXiv:1811.03804v4 [cs.LG]). Retrieved from http://arxiv.org/abs/1811.03804

Dumas, B., & Luciano, E. (1991, 6). An Exact Solution to a Dynamic Portfolio Choice Problem under Transactions Costs. *The Journal of Finance*, *46*(2), 577. Retrieved from https://www.jstor.org/stable/2328837?origin=crossref doi: 10.2307/2328837

Dybvig, P. H. (1984, 3). Short Sales Restrictions and Kinks on the Mean Variance Frontier. *The Journal of Finance*, *39*(1), 239–244. Retrieved from http://doi.wiley.com/10.1111/j.1540-6261.1984.tb03871.x doi: 10.1111/j.1540-6261.1984.tb03871.x

Egmont-Petersen, M., de Ridder, D., & Handels, H. (2002, 10). Image processing with neural networks—a review. *Pattern Recognition*, *35*(10), 2279–2301. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S0031320301001789 doi: 10.1016/S0031-3203(01)00178-9

Elman, J. L. (1990, 6). Finding structure in time. *Cognitive Science*, *14*(2), 179–211. Retrieved from http://doi.wiley.com/10.1016/0364

-0213(90)90002-E doi: 10.1016/0364-0213(90)90002-E

Elton, E. J., & Gruber, M. J. (1974). On the Optimality of Some Multiperiod Portfolio Selection Criteria. *The Journal of Business*, *47*(2), 231–243. doi: 10.1086/295633

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017, 2). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118. Retrieved from http://www.nature.com/articles/nature21056 doi: 10.1038/nature21056

Fama, E. F. (1970, 5). Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance*, *25*(2), 383–417. Retrieved from https://www.jstor.org/stable/2325486?origin=crossref doi: 10.2307/2325486

Fama, E. F., & French, K. R. (1993, 2). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, *33*(1), 3–56. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/0304405X93900235 doi: 10.1016/0304-405X(93)90023-5

Fama, E. F., & French, K. R. (2019). *Kenneth r. french - data library.* Retrieved 2019-08-08, from https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html ([online] Available at: https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html [Accessed 2019-08-08])

Gal, Y., & Ghahramani, Z. (2015, 12). A Theoretically Grounded Application of Dropout in Recurrent Neural Networks. (arXiv:1512.05287v5 [stat.ML]). Retrieved from http://arxiv.org/abs/1512.05287

Gârleanu, N., & Pedersen, L. H. (2018, 8). Efficiently Inefficient Markets for Assets and Asset Management. *The Journal of Finance*, *73*(4), 1663–1712. Retrieved from http://doi.wiley.com/10.1111/jofi.12696 doi: 10.1111/jofi.12696

Gately, E. (1995). *Neural networks for financial forecasting*. New York: John Wiley & Sons.

Gepperth, A., & Hammer, B. (2016). Incremental learning algorithms and applications. *ESANN 2016 - 24th European Symposium on Artificial Neural Networks*(April), 357–368.

Gers, F. A., & Schmidhuber, J. (2000). Recurrent nets that time and count. In *Proceedings of the ieee-inns-enns international joint conference on neural networks. ijcnn 2000. neural computing: New challenges and perspectives for the new millennium* (pp. 189–194). IEEE. Retrieved from http://ieeexplore.ieee.org/document/861302/ doi: 10.1109/IJCNN.2000.861302

Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep sparse rectifier neural networks. In G. Gordon, D. Dunson, & M. Dudík (Eds.), *Proceedings of the fourteenth international conference on artificial intelligence and statistics* (Vol. 15, pp. 315–323). Fort Lauderdale: PMLR. Retrieved from http://proceedings.mlr.press/v15/glorot11a.html

Goetzmann, W. N., & Huang, S. (2018, 12). Momentum in Imperial Russia. *Journal of Financial Economics*, *130*(3), 579–591. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S0304405X18301843 doi: 10.1016/j.jfineco.2018.07.008

Goldberg, Y. (2016, 11). A Primer on Neural Network Models for Natural Language Processing. *Journal of Artificial Intelligence Research*, *57*, 345–420. Retrieved from https://jair.org/index.php/jair/article/view/11030 doi: 10.1613/jair.4992

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. Cambridge: MIT Press.

Graves, A., Liwicki, M., Fernandez, S., Bertolami, R., Bunke, H., & Schmidhuber, J. (2009, 5). A Novel Connectionist System for Unconstrained Handwriting Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(5), 855–868. Retrieved from http://ieeexplore.ieee.org/document/4531750/ doi: 10.1109/TPAMI.2008.137

Graves, A., Mohamed, A.-r., & Hinton, G. (2013, 3). Speech Recognition with Deep Recurrent Neural Networks. (arXiv:1303.5778v1 [cs.NE]). Retrieved from http://arxiv.org/abs/1303.5778

Grefenstette, E., Blunsom, P., de Freitas, N., & Hermann, K. M. (2014, 4). A Deep Architecture for Semantic Parsing. (arXiv:1404.7296v1 [cs.CL]). Retrieved from http://arxiv.org/abs/1404.7296

Hahnloser, R. H. R., Sarpeshkar, R., Mahowald, M. A., Douglas, R. J., & Seung, H. S. (2000, 6). Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*, *405*(6789), 947–951. Retrieved from http://www.nature.com/articles/35016072 doi: 10.1038/35016072

Harvey, D., Leybourne, S., & Newbold, P. (1997, 6). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, *13*(2), 281–291. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S0169207096007194 doi: 10.1016/S0169-2070(96)00719-4

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. New York: Springer. Retrieved from http://link.springer.com/10.1007/978-0-387-84858-7 doi: 10.1007/978-0-387-84858-7

Hayou, S., Doucet, A., & Rousseau, J. (2018, 5). On the Selection of Initialization and Activation Function for Deep Neural Networks. (arXiv:1805.08266v2 [stat.ML]). Retrieved from http://arxiv.org/abs/1805.08266

Hill, T., O'Connor, M., & Remus, W. (1996, 7). Neural Network Models for Time Series Forecasts. *Management Science*, *42*(7), 1082–1092. Retrieved from http://pubsonline.informs.org/doi/abs/10.1287/mnsc.42.7.1082 doi: 10.1287/mnsc.42.7.1082

Hinton, G., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012, 7). Improving neural networks by preventing co-adaptation of feature detectors. (arXiv:1207.0580v1 [cs.NE]). Retrieved from http://arxiv.org/abs/1207.0580

Hochreiter, S. (1991). *Untersuchungen zu dynamischen neuronalen Netzen [Investigations of dynamic neural nets]* (Unpublished doctoral dissertation). Technical University of Munich.

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. IEEE Press.

Hochreiter, S., & Schmidhuber, J. (1997, 11). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. Retrieved from http://www.mitpressjournals.org/doi/10.1162/neco.1997.9.8.1735 doi: 10.1162/neco.1997.9.8.1735

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, *4*(2), 251–257. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/089360809190009T doi: 10.1016/0893-6080(91)90009-T

Hurst, B., Ooi, Y. H., & Pedersen, L. H. (2017). A Century of Evidence on Trend-Following Investing. *SSRN Electronic Journal*. Retrieved from https://www.ssrn.com/abstract=2993026 doi: 10.2139/ssrn.2993026

Ioffe, S., & Szegedy, C. (2015, 2). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. (arXiv:1502.03167v3 [cs.LG]). Retrieved from http://arxiv.org/abs/1502.03167

Jegadeesh, N., & Titman, S. (1993, 3). Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency. *The Journal of Finance*, *48*(1), 65–91. Retrieved from http://doi.wiley.com/10.1111/j.1540-6261.1993.tb04702.x doi: 10.1111/j.1540-6261.1993.tb04702.x

Jones, C. M. (2002). A century of stock market liquidity and trading costs. *SSRN Electronic Journal*.

Jordan, M. I. (1997). Serial Order: A Parallel Distributed Processing Approach. In *Advances in psychology, volume 121* (pp. 471–495). Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S0166411597801112 doi: 10.1016/S0166-4115(97)80111-2

J.P. Morgan. (1996). *J.p. morgan/reuters riskmetricstm- technical document*. New York: J.P. Morgan.

Kaastra, I., & Boyd, M. (1996, 4). Designing a neural network for forecasting financial and economic time series. *Neurocomputing*, *10*(3), 215–236. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/0925231295000399 doi: 10.1016/0925-2312(95)00039-9

Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). A Convolutional Neural Network for Modelling Sentences. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 655–665). Stroudsburg: Asso-

ciation for Computational Linguistics. Retrieved from http://aclweb.org/anthology/P14-1062 doi: 10.3115/v1/P14-1062

Kang, L., Ye, P., Li, Y., & Doermann, D. (2014, 6). Convolutional Neural Networks for No-Reference Image Quality Assessment. In *2014 ieee conference on computer vision and pattern recognition* (pp. 1733–1740). IEEE. Retrieved from http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6909620 doi: 10.1109/CVPR.2014.224

Kaufman, S., Rosset, S., & Perlich, C. (2011). Leakage in data mining. In *Proceedings of the 17th acm sigkdd international conference on knowledge discovery and data mining - kdd '11* (p. 556). New York: ACM Press. Retrieved from http://dl.acm.org/citation.cfm?doid=2020408.2020496 doi: 10.1145/2020408.2020496

Kim, Y. (2014, 8). Convolutional Neural Networks for Sentence Classification. (arXiv:1408.5882v2 [cs.CL]). Retrieved from http://arxiv.org/abs/1408.5882

Kingma, D. P., & Ba, J. (2014, 12). Adam: A Method for Stochastic Optimization. (arXiv:1412.6980v9 [cs.LG]). Retrieved from http://arxiv.org/abs/1412.6980

Krauss, C., Do, X. A., & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research*, *259*(2), 689–702. Retrieved from http://dx.doi.org/10.1016/j.ejor.2016.10.031 doi: 10.1016/j.ejor.2016.10.031

Laptev, N., Yosinski, J., Li, L. E., & Smyl, S. (2017). Time-series extreme event forecasting with neural networks at uber. *International Conference on Machine Learning*, *34*, 1–5.

Lawrence, S., Giles, C., Ah Chung Tsoi, & Back, A. (1997). Face recognition: a convolutional neural-network approach. *IEEE Transactions on Neural Networks*, *8*(1), 98–113. Retrieved from http://ieeexplore.ieee.org/document/554195/ doi: 10.1109/72.554195

LeCun, Y., Bengio, Y., & Hinton, G. (2015, 5). Deep learning. *Nature*, *521*(7553), 436–444. Retrieved from http://www.nature.com/articles/nature14539 doi: 10.1038/nature14539

LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., & Jackel, L. D. (1989, 12). Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, *1*(4), 541–551. Retrieved from http://www.mitpressjournals.org/doi/10.1162/neco.1989.1.4.541 doi: 10.1162/neco.1989.1.4.541

LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., & Jackel, L. D. (1990). Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems* (pp. 396–404).

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. Retrieved from http://ieeexplore.ieee.org/document/726791/ doi: 10.1109/5.726791

Ledoit, O., & Wolf, M. (2003, 12). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, *10*(5), 603–621. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S0927539803000070 doi: 10.1016/S0927-5398(03)00007-0

Ledoit, O., & Wolf, M. (2004, 2). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, *88*(2), 365–411. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S0047259X03000964 doi: 10.1016/S0047-259X(03)00096-4

Ledoit, O., & Wolf, M. (2008, 12). Robust performance hypothesis testing with the Sharpe ratio. *Journal of Empirical Finance*, *15*(5), 850–859. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S0927539808000182 doi: 10.1016/j.jempfin.2008.03.002

Lempérière, Y., Deremble, C., Seager, P., Potters, M., & Bouchaud, J. P. (2014, 4). Two centuries of trend following. (arXiv:1404.3274v1 [q-fin.PM]). Retrieved from http://arxiv.org/abs/1404.3274

Li, D., & Ng, W.-L. (2000, 7). Optimal Dynamic Portfolio Selection: Multiperiod Mean-Variance Formulation. *Mathematical Finance*, *10*(3), 387–406. Retrieved from http://doi.wiley.com/10.1111/1467-9965.00100 doi: 10.1111/1467-9965.00100

Li, F.-F. (2019). *Stanford University CS231n: Optimization.* ([online] Available at: http://cs231n.github.io/optimization-1/ [Accessed 2019-09-08])

Liew, J. K., & Mayster, B. (2017, 12). Forecasting ETFs with Machine Learning Algorithms. *The Journal of Alternative Investments*, *20*(3), 58–78. Retrieved from http://jai.iijournals.com/lookup/doi/10.3905/jai.2018.20.3.058 doi: 10.3905/jai.2018.20.3.058

Lim, B., Zohren, S., & Roberts, S. (2019). Enhancing Time Series Momentum Strategies Using Deep Neural Networks. (arXiv:1904.04912v1 [stat.ML]). Retrieved from http://arxiv.org/abs/1904.04912

Livni, R., Shalev-Shwartz, S., & Shamir, O. (2014, 10). On the Computational Efficiency of Training Neural Networks. (arXiv:1410.1141v2 [cs.LG]). Retrieved from http://arxiv.org/abs/1410.1141

Lopez, J. A., & Walter, C. A. (2002). Evaluating Covariance Matrix Forecasts in a Value-at-Risk Framework. *SSRN Electronic Journal*. Retrieved from http://www.ssrn.com/abstract=305279 doi: 10.2139/ssrn.305279

Markowitz, H. (1952, 3). Portfolio Selection. *The Journal of Finance*, *7*(1), 77. Retrieved from https://www.jstor.org/stable/2975974?origin=crossref doi: 10.2307/2975974

Masters, D., & Luschi, C. (2018, 4). Revisiting Small Batch Training for Deep Neural Networks. (arXiv:1804.07612v1 [cs.LG]). Retrieved from http://arxiv.org/abs/1804.07612

McCulloch, W. S., & Pitts, W. (1943, 12). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133. Retrieved from http://link.springer.com/10.1007/BF02478259 doi: 10.1007/BF02478259

Merton, R. C. (1972, 9). An Analytic Derivation of the Efficient Portfolio Frontier. *The Journal of Financial and Quantitative Analysis*, *7*(4), 1851. Retrieved from https://www.jstor.org/stable/2329621?origin=crossref doi: 10.2307/2329621

Michaud, R. O. (1989, 1). The Markowitz Optimization Enigma: Is 'Optimized' Optimal? *Financial Analysts Journal*, *45*(1), 31–42. Retrieved from https://www.tandfonline.com/doi/full/10.2469/faj.v45.n1.31 doi: 10.2469/faj.v45.n1.31

Müller, A. C., & Guido, S. (2016). *Introduction to Machine Learning with Python*. Sebastopol: OŔeilly Media. Retrieved from https://www.oreilly.com/library/view/introduction-to-machine/9781449369880/

Morantz, B., Whalen, T., & Zhang, G. P. (2008). Neural Network Time Series Forecasting Using Recency Weighting. In *Encyclopedia of decision making and decision support technologies* (pp. 661–667). Hershey: IGI Global. Retrieved from http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-59904-843-7.ch074 doi: 10.4018/978-1-59904-843-7.ch074

Mossin, J. (1968). Optimal Multiperiod Portfolio Policies. *The Journal of Business*, *41*(2), 215–229.

Mozer, M. C. (1989). A Focused Backpropagation Algorithm for Temporal Pattern Recognition. *Complex Systems*, *3*, 349- 381.

Ng, A. (2019a). *Stanford University CS229: Lecture notes.* Retrieved 2019-09-08, from http://cs229.stanford.edu/notes/cs229-notes1.pdf ([online] Available at: http://cs229.stanford.edu/notes/cs229-notes1.pdf [Accessed 2019-09-08])

Ng, A. (2019b). *Stanford University CS229: Splitting into train, dev and test sets.* Retrieved 2019-09-08, from https://cs230-stanford.github.io/train-dev-test-split.html ([online] Available at: https://cs230-stanford.github.io/train-dev-test-split.html [Accessed 2019-09-08])

Nickolls, J., Buck, I., Garland, M., & Skadron, K. (2008, 3). Scalable parallel programming with CUDA. *Queue - GPU Computing*, *6*(2), 40–53. Retrieved from http://portal.acm.org/citation.cfm?doid=1365490.1365500 doi: 10.1145/1365490.1365500

OpenAI. (2018). *OpenAI Five.* Retrieved from https://blog.openai.com/openai-five/ ([online] Available at: https://blog.openai.com/openai-five/ [Accessed 2019-09-08])

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. ([online] Available at: https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf [Accessed 2019-09-08])

Robinson, A. J., & Fallside, F. (1987). *The utility driven dynamic error propagation network*. University of Cambridge, Department of Engineering Cambridge.

Ross, S. A. (1977, 3). The Capital Asset Pricing Model (CAPM), Short-Sale Restrictions and Related Issues. *The Journal of Finance*, *32*(1), 177. Retrieved from https://www.jstor.org/stable/2326912?origin=crossref doi: 10.2307/2326912

Ruder, S. (2016, 9). An overview of gradient descent optimization algorithms. (arXiv:1609.04747v2 [cs.LG]). Retrieved from http://arxiv.org/abs/1609.04747

Samuelson, P. A. (1969, 8). Lifetime Portfolio Selection By Dynamic Stochastic Programming. *The Review of Economics and Statistics*, *51*(3), 239. Retrieved from https://www.jstor.org/stable/1926559?origin=crossref doi: 10.2307/1926559

Schmidhuber, J. (2015, 1). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S0893608014002135 doi: 10.1016/j.neunet.2014.09.003

Severini, T. A. (2017). *Introduction to statistical methods for financial models*. Boca Raton: CRC Press, Taylor & Francis Group.

Sharpe, W. F. (1966, 1). Mutual Fund Performance. *The Journal of Business*, *39*(1), 119. Retrieved from https://www.jstor.org/stable/2351741 doi: 10.1086/294846

Sharpe, W. F. (1994, 10). The Sharpe Ratio. *The Journal of Portfolio Management*, *21*(1), 49–58. Retrieved from http://jpm.pm-research.com/lookup/doi/10.3905/jpm.1994.409501 doi: 10.3905/jpm.1994.409501

Smith, K. V. (1967, 9). A Transition Model for Portfolio Revision. *The Journal of Finance*, *22*(3), 425. Retrieved from https://www.jstor.org/stable/2978895?origin=crossref doi: 10.2307/2978895

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, *15*, 1929–1958. Retrieved from http://jmlr.org/papers/v15/srivastava14a.html

Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017, 7). Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. (arXiv:1707.02968v2 [cs.CV]). Retrieved from http://arxiv.org/abs/1707.02968

Swinkels, L. (2004, 8). Momentum investing: A survey. *Journal of Asset Management*, *5*(2), 120–143. Retrieved from http://link.springer.com/10.1057/palgrave.jam.2240133 doi: 10.1057/palgrave.jam.2240133

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013, 12). Intriguing properties of neural networks. (arXiv:1312.6199v4 [cs.CV]). Retrieved from http://arxiv.org/abs/1312.6199

Theodoridis, S., & Koutroumbas, K. (2009). *Pattern recognition*. Burlington: Academic Press.

Tino, P., Schittenkopf, C., & Dorffner, G. (2001, 7). Financial volatility trading using recurrent neural networks. *IEEE Transactions on Neural Networks*, *12*(4), 865–874. Retrieved from http://ieeexplore.ieee.org/document/935096/ doi: 10.1109/72.935096

Wakker, P. P. (2008, 12). Explaining the characteristics of the power (CRRA) utility family. *Health Economics*, *17*(12), 1329–1344. Retrieved from http://doi.wiley.com/10.1002/hec.1331 doi: 10.1002/hec.1331

Walter, C. A., & Lopez, J. A. (2000, 2). Is Implied Correlation Worth Calculating? *The Journal of Derivatives*, *7*(3), 65–81. Retrieved from http://jod.pm-research.com/lookup/doi/10.3905/jod.2000.319125 doi: 10.3905/jod.2000.319125

Werbos, P. J. (1988, 1). Generalization of backpropagation with application to a recurrent gas market model. *Neural Networks*, *1*(4), 339–356. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/089360808890007X doi: 10.1016/0893-6080(88)90007-X

Würtz, D., Ellis, A., Chalabi, Y., Packages, R., Chalabi, Y., Chen, W., & Ellis, A. (2009). *Portfolio optimization with r/rmetrics rmetrics, association & finance online.* ([online] Available at: https://www.rmetrics.org/downloads/9783906041018-fPortfolio.pdf [Accessed 2019-09-08])

Junior Management Science 6(1) (2021) 190-236

Junior Management Science

journal homepage: www.jums.academy

Analysis of Global Tax Standards of the OECD on Digital Business Models

Analyse globaler Steuerstandards der OECD zu digitalen Geschäftsmodellen

Christina Winder

Universität Liechtenstein

Abstract

With the emergence of new digital business models, the effectiveness of international tax law is increasingly disputed. The discussion whether tax law is still capable of taxing digital business models correctly and fairly is also being further ignited by US corporations. The OECD is working intensively on a global proposal to be able to tax digital corporations now in market states as well. The following research question "Which challenges arise upon the taxation of digital business models and which global tax standards are proposed by the OECD in relation to the taxation of digital business models?" is answered by this thesis.

This bachelor thesis illustrates that the basic concepts of international tax law are no longer able to effectively tax digital business models. The OECD's proposed solution revolutionizes international tax law and increasingly turns away from the concept of physical presence. The proposal will lead to a shift of taxation substrate towards market states. The OECD's Unified Approach is capable of reducing and solving existing problems in the taxation of the digital economy.

Zusammenfassung

Durch das Aufkommen neuer digitaler Geschäftsmodelle wird die Effektivität des internationalen Steuerrechts immer weiter angezweifelt. Die Diskussion, ob das Steuerrecht noch in der Lage ist, die digitalen Geschäftsmodelle richtig und fair zu besteuern, wird auch durch die US-amerikanischen Konzerne weiter entfacht. Die OECD arbeitet intensiv an einem globalen Vorschlag, um Digitalkonzerne nun auch in den Marktstaaten besteuern zu können. Folgende Forschungsfrage „Welche Herausforderungen entstehen bei der Besteuerung von digitalen Geschäftsmodellen und welche globalen Steuerstandards werden von der OECD in Bezug auf die Besteuerung digitaler Geschäftsmodelle vorgeschlagen?" wird mittels dieser Thesis beantwortet.

Diese Bachelorthesis verdeutlicht, dass die Grundkonzepte des internationalen Steuerrechts nicht mehr in der Lage sind, die digitalen Geschäftsmodelle effektiv zu besteuern. Der geplante Lösungsansatz der OECD revolutioniert das internationale Steuerrecht und wendet sich immer mehr vom Konzept der physischen Präsenz ab. Der Vorschlag wird zu einem Verschieben von Besteuerungssubstrat in die Marktstaaten führen. Der Unified Approach der OECD ist durchaus in der Lage, bestehende Probleme bei der Besteuerung der Digitalwirtschaft zu verringern und zu lösen.

Keywords: Digitale Geschäftsmodelle; internationale Unternehmensbesteuerung; Unified Approach; globale Steuerstandards.

1. Einleitung

Die Digitalisierung beeinflusst und verändert nahezu jeden Lebensbereich auf bedeutsame Weise.[1] Schuster hebt hervor, dass die Digitalisierung viele Gesetzgeber und Rechtsbereiche vor neue Herausforderungen und Probleme stellt. Während einige Rechtsbereiche angepasst werden können, benötigt es von anderen Rechtsbereichen ein komplettes Umdenken und das Betreten von Neuland. Besonders davon betroffen ist das Steuerrecht. Die Veränderungen aufgrund der Digitalisierung sind von hoher Geschwindigkeit und Intensität, weshalb die grundlegenden Prinzipien des

[1] Marco Boehle und Tobias Scholtholt, „Gestaltungsoptionen zur gerechten Besteuerung von Digitalkonzernen: Eine kritische Diskussion zur aktuellen Arbeit des BEPS-Projekts der OECD," IStR, Nr. 23 (2019): 919.

DOI: https://doi.org/10.5282/jums/v6i1pp190-236

Steuerrechts neu beurteilt werden müssen.[2] Im Jahr 2014 betont die Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (OECD), dass es zwingend notwendig ist, die bestehenden Steuersysteme zu überdenken. Folglich müssen neue und internationale Standards erlassen werden, um die Besteuerung von Unternehmenseinkünften, die über Landesgrenzen hinaus auftreten, abzustimmen.[3] Aufgrund dieser Notwendigkeit beschäftigt sich die OECD mit der Erarbeitung eines neuen globalen Steuerstandards. Das Projekt der OECD, welches in zwei Säulen aufgeteilt werden kann, soll sich der Besteuerung digitaler Unternehmen widmen. Während sich Säule 1 der Aufteilung von Besteuerungsrechten und Gewinnen zwischen Marktjurisdiktionen widmet, soll im Rahmen der zweiten Säule eine Mindestbesteuerung multinationaler Unternehmen (MNE) eingeführt werden.[4]

Schön beschreibt die Veränderungen in der internationalen Steuerpolitik wie folgt: „Steuerpolitik wird jetzt wahrhaft global betrieben – und bedarf eines globalen, sich entfaltenden Konsenses. Das Jahr 2019 – es könnte zum Schicksalsjahr des internationalen Steuerrechts werden.“ Schön weist mit diesem Zitat darauf hin, dass 2019 die Diskussion zur Besteuerung der digitalen Wirtschaft seinen Höhepunkt erreichte. Diese Diskussion führte im Jahr 2019 nun zu einer grundlegenden und weltweiten Debatte, inwieweit die Besteuerungsrechte zwischen Markt- und Produktionsstaaten aufgeteilt werden können.[5] Auch Bendlinger macht darauf aufmerksam, dass insbesondere im internationalen Steuerrecht die Sicherung von Besteuerungsrechten an der digitalen Wirtschaft zu grossen Diskussionen führt. Durch die bisherigen Base Erosion and Profit Shifting (BEPS)-Arbeiten der OECD ergibt sich zunehmend die Frage, inwieweit in der Zukunft Besteuerungsrechte zwischen Jurisdiktionen verteilt werden sollen.[6]

Versin und Hidien weisen zudem darauf hin, dass die Diskussion rund um die Besteuerung der digitalen Wirtschaft, insbesondere der GAFA-Konzerne[7], ein hoch brisantes politisches Thema darstellt.[8] Laut Zinowsky wollen Länder, in denen Digitalunternehmen beachtliche Gewinne ohne jegliche physische Präsenz erzielen, ein grösseres Stück vom „Steuerkuchen“ abbekommen und bestehen deshalb auf die Anpassung der traditionellen Steuerregeln an die digitale Wirtschaft.[9] Schön zufolge suche die USA bereits seit Längerem eine Möglichkeit, die Benachteiligung, dass auf Importe aus den USA Umsatzsteuer erhoben wird, während Importe in die Vereinigten Staaten keiner direkten oder indirekten Steuer unterliegen, zu kompensieren. Dabei wird die USA von Entwicklungs- und Schwellenländern unterstützt, da sich diese sowohl als Kundenstaaten wie auch als kostengünstige Produktionsstaaten betrachten und somit fiskalpolitische Vorteile daraus ziehen können. Für die exportorientierten europäischen Länder wird es bedeutend schwerer, sich gegen eine Allianz - mit Hauptbefürworter USA - langfristig zu behaupten.[10] Auch Duttiné macht darauf aufmerksam, dass die Modernisierung des internationalen Steuerrechts nicht den politischen Machtspielen zum Opfer fallen soll, sondern eine nachhaltige, verlässliche und faire Lösung herausgearbeitet werden muss.[11]

Die untenstehende Tabelle 1 zeigt, dass die Umsätze, welche von Gewinnverschiebungs- und Gewinnverlagerungsmassnahmen betroffen sind, auf rund 650 Mrd. US-Dollar quantifiziert werden können. Die grössten Gewinner durch Gewinnverschiebungsmassnahmen der Unternehmen sind Irland, die Karibik, die Niederlande, Singapur und die Schweiz. Die grössten Verlierer der Gewinnverschiebungsmassnahmen der MNE sind die USA, Grossbritannien sowie auch Deutschland, China und Frankreich.[12] In einer vorläufigen Analyse und Folgenabschätzung zu Säule 1 verdeutlichen die Daten der OECD zudem, dass die Länder mit Investitionsdrehscheiben die Verlierer bezüglich Steuereinnahmen sein werden.[13]

Obwohl sich die Länder uneinig bezüglich einer Besteuerung digitaler Geschäftsmodelle sind, umso geschlossener vertreten sie die Meinung, dass eine Veränderung des internationalen Steuerrechts unabdingbar ist. Es herrscht einheitlicher Konsens darüber, dass das momentane Steuerrecht nicht mehr an die Weltwirtschaft angepasst werden kann und das Konzept der physischen Präsenz ausgedient hat und somit erweitert werden muss.[15] Die Relevanz des Themas scheint unbestreitbar, da die Einflussnahme der OECD im internationalen Steuerrecht hoch ist. In einem Ernst & Young-Webcast im Jahre 2019 gaben 75 % der Befragten an, dass die Einführung des geplanten Zwei-Säulen-Ansatzes der OECD einen starken bzw. wesentlichen Einfluss auf ihr Unternehmen haben wird. Da die geplanten Umsetzungen der OECD nicht nur Unternehmen der digitalen Wirtschaft betreffen, sondern weit darüber hinaus gehen, ist es für internationale Unternehmungen von grosser Bedeutung, die

[2] Joachim Schuster, „Herkulesaufgabe Digitalisierung: Herausforderungen an das EU-Recht: Gastkommentar,“ EuZW 30, Nr. 24 (2019): 1001-2002.

[3] OECD, Aktionsplan zur Bekämpfung der Gewinnverkürzung und Gewinnverlagerung (Paris: OECD Publishing, 2014), 7-21.

[4] OECD, „Addressing the Tax Challenges of the Digitalisation of the Economy - Policy Note,“ 23 January 2019, OECD/G20 Base Erosion and Profit Shifting Project, 1-2.

[5] Wolfgang Schön, „Seminar F (IFA/OECD): Internationale Besteuerung der digitalisierten Wirtschaft,“ IStR, Nr. 16 (2019): 650.

[6] Stefan Bendlinger, „BEPS 2.0 - Besteuerung digitalisierter Geschäfte - eine (un)endliche Geschichte,“ WT Fachjournal, Nr. 02 (2019): 108-110.

[7] Google, Apple, Facebook und Amazon

[8] Volker Versin und Jürgen W. Hidien, „Digitales Geschäftsmodell als steuerlicher Wettbewerbsfaktor,“ GmbHR, Nr. 14 (2019): 761.

[9] Tim Zinowsky, „Abkehr von tradierten Verrechnungspreismethoden und Einführung einer weltweiten Mindestbesteuerung: Diskussion des Arbeitsprogramms für die Verabschiedung einer neuen “Weltsteuerordnung”,“ IStR, Nr. 20 (2019): 812.

[10] Schön, „Seminar F (IFA/OECD): Internationale Besteuerung der digitalisierten Wirtschaft,“ 648.

[11] Tino Duttiné, „Lieber gut als schnell. Keine populistischen Lösungen für die Besteuerung der digitalen Wirtschaft,“ BB, Nr. 48 (2019): 1.

[12] Thomas Tørsløv, Ludvig Wier und Gabriel Zucman, „40% of multinational profits are shifted to tax havens each year,“ Zugriff am 03.04.2020, https://missingprofits.world/.

[13] OECD, „Tax Challenges arising from the Digitalisation of the Economy: Update on the Economic Analysis & Impact Assessment,“ Webcast 13 February 2020, 12.

[14] Tørsløv, Wier und Zucman, „40% of multinational profits are shifted to tax havens each year.“

[15] Lilian V. Faulhaber, „Taxing Tech: The Future of Digital Taxation,“ Virginia Tax Review, Nr. 145 (2019): 51-52.

Tabelle 1: Gewinner und Verlierer von Gewinnverschiebungsmassnahmen

Eigene Darstellung in Anlehnung an Tørsløv, Wier und Zucman[14]

Gewonnene Umsätze der Länder aufgrund Gewinnverschiebungen in Mrd. USD		Verlorene Umsätze der Länder aufgrund Gewinnverschiebungen in Mrd. USD	
1 Irland	117,11	1 USA	152,23
2 Karibik	92,83	2 Grossbritannien	80,61
3 Niederlande	90,26	3 Deutschland	65,46
4 Singapur	77,59	4 China	50,48
5 Schweiz	73,17	5 Frankreich	36,04

zukünftigen Entwicklungen und Veränderungen im internationalen Steuerrecht zu verfolgen.[16]

1.1. Ausgangssituation und Entstehung des Zwei-Säulen-Ansatzes

Boehle und Scholtholt erwähnen, dass bereits am Ende des 20. Jahrhunderts die Erfassung und Besteuerung von digitalen Geschäftsmodellen diskutiert wurde. Schon damalige Diskussionen erkennen die Hauptproblematik des Steuerrechts, welches mit dem Zitat „Business goes global, taxes stay local“ treffend beschrieben wird.[17] Bereits 1988 beschäftigen sich die Mitgliedstaaten der OECD beim Forum on Electronic-Commerce in Ottawa mit der Frage, wie mit den neu auftretenden Geschäftsmodellen steuerrechtlich umgegangen werden soll. Es wurde eine Einigung dahingehend erzielt, dass steuerrechtlich kein Unterschied bezüglich unternehmerischer Aktivitäten von digitalen und konventionellen Geschäftsmodellen bestehen darf.[18]

Im Jahr 2013 wird die OECD von den G20-Staaten beauftragt, ein Lösungskonzept zur Eindämmung der Gewinnverlagerungen und Gewinnvermeidungsstrategien digitaler Konzerne zu erarbeiten. Dieser Auftrag wurde insbesondere von europäischen Mitgliedstaaten gefordert, da viele amerikanische Technologie-Konzerne ihre Steuerlast in Europa durch bewusste Konzernstrukturen auf ein Minimum reduzieren können oder gar keine Steuerlast aufweisen. Dies kann durch die lückenhaften CFC-Rules[19] des US-Fiskus erklärt werden. Die Tatsache, dass sich von den 60 wertvollsten digitalen Plattformen rund 67 % in den USA befinden, verdeutlicht, dass diese Gewinnverlagerungs- und Vermeidungsstrategien der digitalisierten Unternehmen sehr wohl bekannt sind, der US-Fiskus diese Lücken jedoch nicht schliessen will, um die steuerlichen Wettbewerbsvorteile zu behalten.[20]

Die OECD befasst sich seit einigen Jahren mit der Globalisierung und Digitalisierung sowie den daraus entstehenden Auswirkungen auf die bestehenden Steuersysteme. Steuerpolitik kann nicht mehr nur isoliert und national abgewickelt werden, sondern kann inzwischen, hervorgerufen durch Globalisierung und Digitalisierung, viele verschiedene nationale Gesetzgebungen betreffen. Die Entstehung zahlreicher digitalisierter Unternehmungen führt dazu, dass immer mehr Schwächen und Lücken in den bisherigen Steuersystemen ausgenutzt werden können. Hierbei geht es vor allem um Fälle, die von doppelter Nichtbesteuerung oder Niedrigbesteuerung betroffen sind.[21] Für Staaten ist es von grosser Bedeutung, gegen Gewinnverlagerungen und Gewinnverkürzungen der Unternehmen anzukämpfen. Aus diesem Grund beschliessen die G20-Staaten und die OECD 2013 einen Aktionsplan mit 15 Punkten, welcher eine Besteuerung am Ort der Wertschöpfung garantieren und gegen Gewinnverlagerungen und Gewinnverkürzungen Bestand haben soll. Da eine zeitnahe Lösung notwendig ist, wurde das BEPS Massnahmenpaket im Jahr 2015, teilweise auch schon im Jahr 2014, vorgelegt. Der Aktionspunkt 1, unter dem Namen „Herausforderungen für die Besteuerung der digitalen Wirtschaft“, beschäftigt sich mit den Problemen, insbesondere mit den Gewinnverlagerungs- und Gewinnvermeidungsmassnahmen, die sich durch die Digitalisierung und die daraus entstehende digitale Wirtschaft eröffnen. Der Abschlussbericht zum Aktionspunkt 1 wird 2015 veröffentlicht. Die OECD hebt hervor, dass gemeinsame und globale Massnahmen gegen Gewinnverkürzungen und Gewinnverlagerungen notwendig sind und somit auch Zeit und Ressourcen in die Lösungserarbeitung investiert werden müssen.[22] Im Abschlussbericht zum Aktionspunkt 1 werden sowohl die digitale Wirtschaft als auch die neu entstehenden Geschäftsmodelle der digitalen Wirtschaft analysiert. Zudem werden die wesentlichen Merkmale der digitalen Geschäftsmodelle analysiert, um möglichst genauen Einblick zu erhalten, wie im Zeitalter der Digitalisierung Wert geschöpft werden kann. Die OECD stellt sich hierbei auch die Frage, ob die bestehenden Besteuerungsrechte immer noch zeitgemäss

[16] Fabian Berr, Céline Bulard und Nate Zahnd, „Besteuerung der digitalisierten Wirtschaft,“ Expert Focus, Nr. 9 (2019): 754-755.

[17] Boehle und Scholtholt, „Gestaltungsoptionen zur gerechten Besteuerung von Digitalkonzernen,“ 919.

[18] OECD, A Global Action Plan for Electronic Commerce: Prepared by Business with Recommendations and for Govern-ments, OECD Digital Economy Papers, Nr. 44 (Paris: OECD Publishing, 1999), 38.

[19] Controlled Foreign Corporation Rules: Hinzurechnungsbesteuerung

[20] Boehle und Scholtholt, „Gestaltungsoptionen zur gerechten Besteuerung von Digitalkonzernen,“ 920.

[21] OECD, Aktionsplan zur Bekämpfung der Gewinnverkürzung und Gewinnverlagerung, 7-21.

[22] OECD, Herausforderungen für die Besteuerung der digitalen Wirtschaft (Paris: OECD Publishing, 2015), 3-4.

sind und weiter verfolgt werden können. Weiters wird im Endbericht des Aktionspunktes 1 darauf hingewiesen, dass Gewinnverlagerungs- und Gewinnvermeidungsmassnahmen nicht nur aufgrund der digitalen Wirtschaft bestehen, sondern vor allem durch die digitale Wirtschaft verstärkt werden.[23]

Die Task-Force Digitale Wirtschaft (Taskforce on the Digital Economy – TFDE) wird im Rahmen des Inclusive Framework on BEPS beauftragt, ihre Arbeit fortzusetzen und einen Zwischenbericht im Jahr 2018 sowie einen Endbericht im Jahr 2020 vorzulegen. Dieser Zwischenbericht, auch Interim Report genannt, wird dazu genutzt, um noch näher zu erörtern, wie sich die Digitalisierung auf Geschäftsmodelle und die Wertschöpfung auswirkt. Dieser Zwischenbericht erweitert somit die Ergebnisse aus dem oben genannten Abschlussbericht des BEPS-Aktionspunktes 1. Die OECD unterstreicht, dass „Steuersysteme für den digitalisierungsbedingten Wandel gerüstet sind, dass die Chancen der Digitalisierung genutzt werden und dass die damit einhergehenden Risiken eingedämmt werden".[24]

Im Jänner 2019 wird vom Inclusive Framework on BEPS eine Policy Note[25] veröffentlicht, in welcher das Zwei-Säulen-Programm zum ersten Mal erwähnt wird und auch die beiden Säulen kurz vorgestellt werden. Während sich Säule 1 mit den umfassenden Herausforderungen der digitalen Wirtschaft beschäftigt, widmet sich Säule 2 einer globalen Mindestbesteuerung. Aufgrund des beschränkten Umfangs dieser Thesis wird im weiteren Verlauf dieser Arbeit lediglich auf Säule 1 näher eingegangen und Säule 2 zur Gänze unberücksichtigt bleiben. Die OECD macht in dieser Policy Note vor allem darauf aufmerksam, dass digitale Unternehmen zusätzliche Probleme schaffen, die über das BEPS-Problem hinausragen. Diesbezüglich wird vor allem das Problem genannt, dass bei einem Einkommen aus grenzüberschreitenden Tätigkeiten unklar ist, wie die Besteuerungsrechte unter den verschiedenen Jurisdiktionen fair aufgeteilt werden können. Schon in der Policy Note wird das Ziel genannt, mehr Besteuerungsrechte in den Markt- bzw. Nutzerstaat zu verlegen. Die anvisierten Vorschläge werden sich somit zunehmend vom Konzept der physischen Präsenz entfernen.

Im Februar 2019 folgt ein Konsultationsdokument, das „Addressing the Tax Challenges of the Digitalisation of the Economy: Public Consultation Document"[26], in welchem drei verschiedene Lösungsvorschläge zur Begründung eines Nexus, also dem steuerlichen Anknüpfungspunkt, präsentiert werden. Diese Vorschläge begründen einen Nexus aufgrund von Nutzerbeteiligung („user participation"), Marketing- und Distributionsfunktionen („marketing intangibles") sowie einer signifikanten digitalen Präsenz („significant economic presence"). Diese Vorschläge dienen als Analyse- und Diskussionsgrundlage und sollen im Zuge dieses Projektes weiterentwickelt werden.[27] Es wird darauf aufmerksam gemacht, dass Lösungen bezüglich des Nexus-Ansatzes eng verbunden mit dazugehörigen Gewinnverteilungsregeln sind, da ohne passende Gewinnverteilungsregeln auch der Nexus-Ansatz nicht funktioniert.[28] Anschliessend wird im Mai 2019 das Programme of Work veröffentlicht, welches die Aussagen aus dem Konsultationspapier als roter Faden weiterführt.[29]

Im Oktober 2019 wird der Unified Approach[30] der OECD vorgestellt, welcher die drei zuvor genannten Vorschläge zur Begründung eines Nexus[31] zu einem Vorschlag zusammenfasst. Die Ausarbeitung erfolgte durch das OECD-Sekretariat. Der Unified Approach soll somit die Vorteile dieser drei Vorschläge verbinden und „zum kleinsten gemeinsamen Nenner zusammenführen", so Schuster.[32] Der Unified Approach beinhaltet weitere Fragen und bildet wiederum eine Diskussionsgrundlage. Zu dieser Diskussionsrunde langen rund 300 Kommentare von Unternehmen und Privatpersonen ein. Diese rege Beteiligung aus den Gebieten der Beratung, Wissenschaft, Wirtschaft und Zivilgesellschaft verdeutlicht die hohe Relevanz dieses Themas.[33] Der Unified Approach wurde mithilfe der eingereichten Kommentare und Anregungen zum „Outline of the Architecture of a Unified Approach on Pillar One"[34] weiterentwickelt und konkretisiert und stellt somit einen fortgeschrittenen Entwurf des Unified Approach dar. Dieses Dokument bildet die Grundlage dieser Thesis und wird unter Kapitel 3 und 4 näher erläutert sowie kritisch analysiert.

In der untenstehenden Abbildung 1 werden die Hauptschritte, ausgehend vom BEPS-Projekt bis zum Outline of the Architecture of a Unified Approach on Pillar One, grafisch dargestellt.

1.2. Forschungsfrage und Zielsetzung

Aufgrund der Aktualität sowie der hohen politischen und internationalen Relevanz der Besteuerung der digitalen Wirtschaft wird sich diese Bachelorarbeit mit der folgenden Forschungsfrage beschäftigen: Welche Herausforderungen entstehen bei der Besteuerung von digitalen Geschäftsmodellen

[23] Ebd., 183-184.

[24] OECD, Steuerliche Herausforderungen der Digitalisierung - Zwischenbericht 2018: Inclusive Framework on BEPS: OECD/G20 Projekt Gewinnverkürzung und Gewinnverlagerung (Paris: OECD Publishing, 2018), 208-209.

[25] OECD, „Addressing the Tax Challenges of the Digitalisation of the Economy - Policy Note," 2-3.

[26] OECD, „Addressing the Tax Challenges of the Digitalisation of the Economy: Public Consultation Document," 13 Feb-ruary - 6 March 2019, OECD/G20 Base Erosion and Profit Shifting Project.

[27] Stefan Schuster, „Der "Unified Approach" der OECD - ein Überblick und erste Einschätzungen," SWI 29, Nr. 11 (2019): 514.

[28] OECD, „Addressing the Tax Challenges of the Digitalisation of the Economy," 8.

[29] Hans Weggenmann, Alexander Blank und Andreas Brunnhübner, „OECD Public Consultation Document/Programme of Work betreffend Vorschläge zur Besteuerung der Digitalwirtschaft - schöne neue Steuerwelt?," IStR 28, Nr. 19 (2019): 769.

[30] OECD, „Secretariat Proposal for a "Unified Approach" under Pillar One: Public Consultation Document," 9 October 2019 - 12 November 2019.

[31] „user participation", „marketing intangibles" und „significant economic presence"

[32] Schuster, „Der "Unified Approach" der OECD - ein Überblick und erste Einschätzungen," 514.

[33] Schön, „Seminar F (IFA/OECD): Internationale Besteuerung der digitalisierten Wirtschaft," 647.

[34] OECD, „Outline of the Architecture of a Unified Approach on Pillar One," 29-30 January 2020, OECD/G20 Base Ero-sion and Profit Shifting Project.

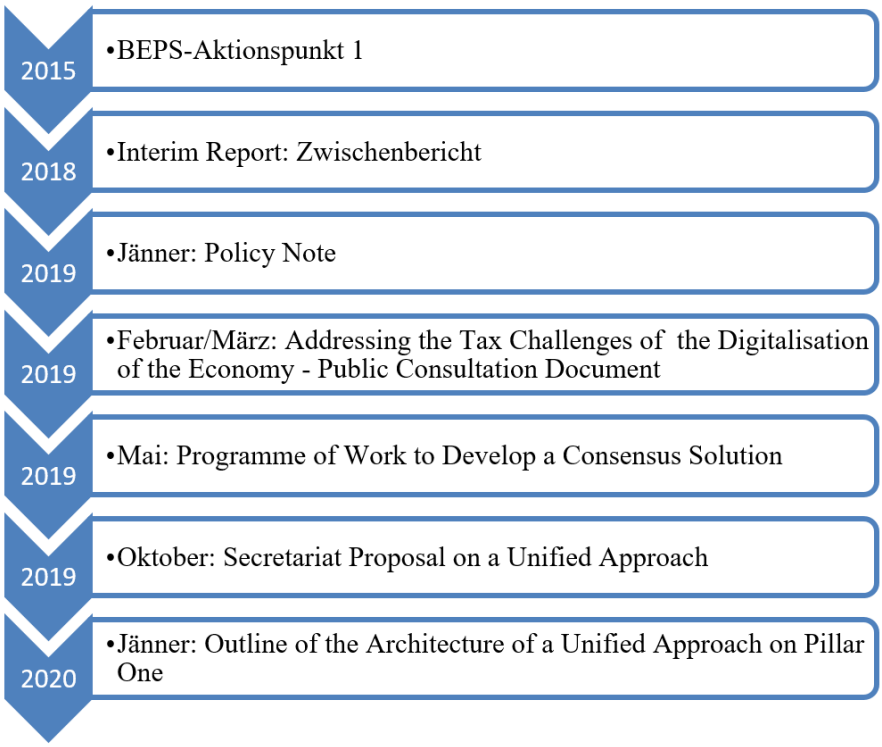


Abbildung 1: Entwicklung BEPS – Outline of the Architecture of a Unified Approach
Eigene Abbildung

und welche globalen Steuerstandards werden von der OECD in Bezug auf die Besteuerung digitaler Geschäftsmodelle vorgeschlagen?

Das Ziel dieser Arbeit besteht darin, einerseits die Herausforderungen, welche bei der Besteuerung von digitalen Geschäftsmodellen auftreten, zu analysieren. Andererseits soll diese Thesis einen Überblick über den aktuellen Stand der Säule 1 des Zwei-Säulen-Ansatzes der OECD geben und deren Einfluss und Auswirkungen auf die zukünftige Besteuerung der digitalen Geschäftsmodelle näher erörtern. Insbesondere soll die Auseinandersetzung mit diesem komplexen Thema zeigen, wie die OECD die Intention, digitalisierte Geschäftsmodelle nun global effektiv zu besteuern, umsetzen will. Ein weiteres Ziel im Zuge dieser Arbeit ist eine Beurteilung des Unified Approach, inwieweit dieser in der Lage ist, die digitalen Geschäftsmodelle nun effektiv zu besteuern und vorhandene Besteuerungsprobleme zu lösen. Diese Bachelorthesis soll Unternehmen sowie auch Steuerberatern und Treuhändern einen Überblick über den aktuellen Stand der Säule 1 geben und zukünftige Auswirkungen darstellen. Aufgrund der weitreichenden Folgen werden in der Zukunft zahlreiche Unternehmungen von dieser Reform des internationalen Steuerrechts betroffen sein. Für diese ist es von wesentlicher Bedeutung, Kenntnisse über die neuesten Entwicklungen auf OECD-Ebene aufzuweisen, da dies sowohl für die Ausübung zukünftiger Geschäftstätigkeiten als auch für in weiterer Folge stattfindende Strategieüberlegungen von Relevanz ist.

1.3. Methodik und Gang der Untersuchung

Um die oben angeführte Forschungsfrage beantworten zu können, bedarf es einer fundierten und ausführlichen Literaturanalyse. Dabei werden die Veröffentlichungen der OECD sowie Journal-Artikel und Aufsätze aus wissenschaftlichen Zeitschriften näher betrachtet und miteinbezogen. Zudem sollen durch Fallbeispiele die ausgeführten Inhalte praxisnah näher gebracht werden. Diese Fallbeispiele werden jedoch stark abstrahiert und vereinfacht dargestellt, um zu gewährleisten, dass die relevanten Inhalte der Fälle möglichst prägnant und klar veranschaulicht werden können.

Das zweite Kapitel dieser Thesis soll die Thematik der digitalen Geschäftsmodelle näher aufgreifen. Um für die folgenden Kapitel Klarheit zu schaffen, werden die Eigenschaften sowie Beispiele digitaler Geschäftsmodelle analysiert. Das Hauptaugenmerk liegt in diesem Kapitel auf der Analyse der Probleme und Herausforderungen, welche bei der Besteuerung digitaler Geschäftsmodelle auftreten. Dieses Kapitel soll als Grundlage für die weiteren Analyseschritte dienen.

Gegenstand des dritten Kapitels soll eine Analyse der Säule 1 des Zwei-Säulen-Ansatzes der OECD sein. Hierbei wird als Grundlage der im Jänner 2020 veröffentlichte Outline of the Architecture of a Unified Approach on Pillar One herangezogen. Die Ziele dieser neuen Steuerordnung sollen genauso wie die neuen und überarbeiteten Nexus- und Gewinnverteilungsregeln näher erläutert und kritisch analysiert werden. Zudem soll auf den Umfang dieser Säule 1 bzw. auf die mög-

lichen Carve-Outs näher eingegangen werden. Der Schwerpunkt dieses dritten Kapitels liegt bei der Analyse des neuen Nexus und den dazugehörenden Gewinnverteilungsregeln unter Betrag A, da diese revolutionierend für das internationale Steuerrecht sind.

Das vierte Kapitel beschäftigt sich mit dem Einfluss der Säule 1 auf die Besteuerung der digitalen Geschäftsmodelle. Es soll mithilfe eines Fallbeispieles die Auswirkungen auf die Besteuerung der digitalen Geschäftsmodelle, welche sich durch Säule 1 ergeben, analysiert werden. Das Ziel dieses Abschnittes ist es, die von der OECD vorgeschlagenen globalen Steuerstandards zu beurteilen und zu erörtern. Der Frage, inwieweit der Unified Approach die vorhandenen Probleme beseitigen oder verringern kann, soll zudem ein Stück weit auf den Grund gegangen werden. Im Anschluss an die Fallbearbeitung folgt eine abschliessende Bewertung des Unified Approach.

1.4. Definitionen

Im folgenden Abschnitt werden nun für das weitere Verständnis relevante Begriffe näher definiert. Zu diesen zählen zweifelsohne die Begriffe Digitalisierung und digitale Geschäftsmodelle sowie die Definition der globalen Steuerstandards.

Digitalisierung kann laut Vornholz mehrere Bedeutungen haben, da es keine allgemeingültige Definition gibt. Der Begriff Digitalisierung beschreibt vor allem die Veränderungen, welche sich durch neue Technologien wie etwa die veränderte Informations- und Kommunikationstechnologie (IKT), durch digitale Geräte und durch das weltumspannende Internet ergeben. Zudem zeichnet sich Digitalisierung oft durch stark vernetzte und intelligente Systeme aus. Die Digitalisierung kann aber auch als technische Transformation verstanden werden. Während die IKT anfangs für eine Modernisierung des Arbeitsplatzes oder zur Einführung von Softwareprogrammen genutzt wurde, liegt das Hauptaugenmerk im 21. Jahrhundert insbesondere auf disruptiven Technologien und innovativen Geschäftsmodellen.[35]

Laut OECD wird die IKT heutzutage von fast allen Unternehmen genutzt, um Flexibilität, Effektivität und die Reichweite der Unternehmen zu verbessern. Sie beeinflusst somit alle bestehenden Geschäftsmodelle.[36] Der Kern digitaler Geschäftsmodelle unterscheidet sich nicht bedeutend von den traditionellen Geschäftsmodellen. Viele digitale Geschäftsmodelle verfolgen die Intention, eine Effizienzsteigerung zu erreichen. Dies wird durch die Verwendung von Daten in Bereichen wie Einkauf, Logistik und Marketing ermöglicht.[37] Die Digitalisierung ist fester Bestandteil des modernen Wirtschaftens geworden und beeinflusst Geschäftsmodelle in unterschiedlichem Ausmass. Es gibt sowohl rein digitale Leistungen (Suchmaschinen, soziale Netzwerke), Kombinationen aus physischen und digitalen Angeboten (Online-Handel) sowie auch klassische Güter, die durch den Fortschritt der Digitalisierung internetgestützt sind (Internet of Things).[38] Versin und Hidien machen darauf aufmerksam, dass es inzwischen zahlreiche rein digitale Geschäftsmodelle gibt. Diese zeichnen sich oft dadurch aus, dass dabei Dienstleistungen im Internet oder in einem elektronischen Netz verkauft werden. Zudem basieren diese Dienstleistungen meist auf der IKT, sind völlig automatisiert und können oft ohne menschliche Unterstützung erfolgen.[39]

Die Definition der globalen Steuerstandards hat sich über die Zeit wesentlich verändert und wurde an verschiedene Bedingungen angepasst. Die Anfänge der globalen Steuerstandards der OECD und EU orientieren sich vor allem aufgrund der Globalisierung daran, dass sowohl international und europaweit Chancengleichheit zwischen den Steuersystemen besteht. Ab 1998 konzentrieren sich diese globalen Steuerstandards dann zunehmend auf die Vermeidung von schädlichen Steuerpraktiken. Zwischen 2009 und 2013 wird einerseits das Ziel verfolgt, Steuertransparenz zu garantieren, während andererseits auch das Prinzip der Besteuerung am Ansässigkeitsort durchgesetzt wird. Daraufhin folgt in den Jahren 2014 bis 2019 das BEPS-Projekt der OECD, welches die Möglichkeiten der MNE, ihre Gewinne in Steueroasen zu verschieben, gezielt einschränken soll. Angelangt im Jahr 2019 wird klar, dass diese globalen Steuerstandards nicht mehr in der Lage sind, neue Geschäftsmodelle fair und effektiv zu besteuern. Als Lösung soll im Jahr 2020 das Zwei-Säulen Projekt der OECD als neuer globaler Steuerstandard folgen und somit eine neue Weltsteuerordnung etablieren.[40]

2. Digitale Geschäftsmodelle

Unternehmen werden inzwischen gezwungen, neue und digitale Wege einzuschlagen, um mit ihren Geschäftsmodellen weiterhin eine Rentabilität zu gewährleisten. Durch die Digitalisierung entstehen aber auch unendlich viele Möglichkeiten, wie die Schaffung gänzlich neuer und innovativer Geschäftsmodelle.[41] Die OECD weist darauf hin, dass durch die Geschwindigkeit der Digitalisierung sehr schnell neue Geschäftsmodelle entstehen können und alte Geschäftsmodelle unbrauchbar werden.[42] Im nachfolgenden Kapitel werden Eigenschaften sowie auch Beispiele digitaler Geschäftsmodelle näher betrachtet. Zudem folgt eine Analyse der Probleme und Herausforderungen, welche durch die digitalen Geschäftsmodelle bei der Besteuerung auftreten.

[35] Günther Vornholz, Digitalisierung der Immobilienwirtschaft (Berlin: De Gruyter, 2019), 25-27.

[36] OECD, Herausforderungen für die Besteuerung der digitalen Wirtschaft, 82.

[37] Versin und Hidien, „Digitales Geschäftsmodell als steuerlicher Wettbewerbsfaktor," 760.

[38] Schön, „Seminar F (IFA/OECD): Internationale Besteuerung der digitalisierten Wirtschaft," 647.

[39] Versin und Hidien, „Digitales Geschäftsmodell als steuerlicher Wettbewerbsfaktor," 760.

[40] Martin Wenz, „The Future of International Taxation: Digitalization and Minimum Taxation - Behavioural Responses - Cases - Lobbying on BEPS Project," February 2020, University of Liechtenstein, 10.

[41] Boehle und Scholtholt, „Gestaltungsoptionen zur gerechten Besteuerung von Digitalkonzernen," 919.

[42] OECD, Herausforderungen für die Besteuerung der digitalen Wirtschaft, 82-85.

2.1. Eigenschaften digitaler Geschäftsmodelle

Die OECD hat durch die Analyse digitaler Geschäftsmodelle drei wesentliche Eigenschaften von digitalen Geschäftsmodellen festgestellt. Dies sind die grenzüberschreitende Ausdehnung ohne Masse („scale without mass"), die Abhängigkeit von immateriellen Werten, insbesondere geistigem Eigentum („reliance on intangible assets") und Daten, Nutzerbeteiligung und entsprechende Synergien mit geistigem Eigentum („data and user participation").[43]

2.1.1. Ausdehnung ohne Masse

Die Digitalisierung ermöglicht vielen Unternehmen, Teile des Produktionsprozesses in verschiedene Staaten auszulagern und zudem auch Kunden aus aller Welt zu generieren. Weiters bietet die Digitalisierung die Möglichkeit, sich stark am Wirtschaftsleben eines Staates zu beteiligen, ohne eine physische Präsenz aufzuweisen. Die Geschäftstätigkeit kann nun also in Orte und Staaten ausgedehnt werden, ohne dort im Besitz von Masse zu sein.[44] Die „Vor-Ort-Präsenz" wird bei digitalen Produkten auf ein Kleinstes minimiert, da die Relevanz des Kontaktes zwischen den Unternehmen und Kunden immer weiter abnimmt.[45]

2.1.2. Abhängigkeit von immateriellen Werten, insbesondere geistigem Eigentum

Eine wesentliche Eigenschaft, die digitale Unternehmen aufweisen, ist die zunehmende Bedeutung der Investitionen in immaterielle Werte, insbesondere in geistiges Eigentum. Das geistige Eigentum kann sich entweder im Besitz des Unternehmens befinden oder auch durch fremde Dritte bereitgestellt werden. Immer mehr Geschäftsmodelle bauen auf der Nutzung von immateriellen Werten und Gegenständen auf, da diese die entscheidenden Funktionen der digitalen Geschäftsmodelle bereitstellen. Beispiele dafür sind Algorithmen und Software, welche als Grundlage für Websites und Plattformen agieren.[46]

2.1.3. Daten und Nutzerbeteiligung

Weitere Eigenschaften der digitalen Geschäftsmodelle sind Daten, Nutzerbeteiligung, Netzwerkeffekte sowie die Bereitstellung nutzergenerierter Inhalte. Die zunehmende Bedeutung der Daten lässt sich auch am Geschäftsmodell der sozialen Netzwerke erkennen, da diese ohne Daten, Nutzerbeteiligung und Netzwerkeffekte in dieser Form, wie sie es heute tun, nicht bestehen und funktionieren würden. Zudem kann zwischen einer aktiven und passiven Nutzerbeteiligung unterschieden werden. Diese hängt jedoch nicht vom Grad der Digitalisierung des Geschäftsmodelles ab. Als Beispiel dafür können Cloud-Dienste genannt werden, die einerseits einen hohen Digitalisierungsgrad aufweisen, jedoch nur wenig Nutzerbeteiligung erfordern.[47]

2.2. Beispiele digitaler Geschäftsmodelle

Die OECD nennt in ihrem Abschlussbericht zum Aktionspunkt 1 verschiedene digitale Geschäftsmodelle wie beispielsweise App-Stores, Online-Werbung, Cloud Computing, Partizipative Netzwerkplattformen, E-Commerce, Hochgeschwindigkeitshandel und Online-Zahlungsdienste. Auf diese Beispiele wird im nachfolgenden Kapitel kurz näher eingegangen. Die OECD hebt hervor, dass sich die Geschäftsmodelle überschneiden bzw. ergänzen können und dies nur eine Auswahl typischer digitaler Geschäftsmodelle ist, jedoch keine endgültige Aufzählung.[48]

2.2.1. App-Stores

App-Stores sind vor allem durch die zunehmende Internetnutzung über Smartphones und Tablets und der damit einhergehenden Nachfrage nach Online-Diensten entstanden. Diese werden oft als Bestandteil des Betriebssystems verkauft und können als eine Vertriebsplattform für Softwareprodukte angesehen werden. Sie fungieren als zentrale Einkaufsplattform und werden für den Nutzer über das eigene Endgerät bereitgestellt. Mithilfe von App-Stores können Apps gefunden, Informationen und Bewertungen eingeholt, Apps gekauft und automatisch am Endgerät des Nutzers installiert werden.[49]

App-Stores offerieren meist sowohl eigene Apps, also vom Betreiber des App-Stores, als auch von Dritten entwickelte Apps. Apps können sowohl gegen Bezahlung als auch gratis verfügbar sein, wobei Letzteres mit hoher Wahrscheinlichkeit durch Werbung finanziert wird.[50] Weiters entstehen immer mehr Apps auf Basis des Freemium-Modells. Hierbei werden die Grundfunktionen gratis zur Verfügung gestellt, zusätzliche Funktionen müssen jedoch käuflich erworben werden. Die Nutzung der App-Stores wird mit zunehmender Digitalisierung stark ansteigen.[51]

2.2.2. Online-Werbung

Die Online-Werbung nutzt das Internet als Plattform, um Werbebotschaften gezielt an Kundenzielgruppen zu vermarkten und zu vertreiben. Ein wesentlicher Vorteil der Online-Werbung ist die Genauigkeit beim Erreichen einer bestimmten Zielgruppe. Viele Anbieter der Online-Werbung verfügen über Methoden, welche die Segmentierung von Kunden ermöglicht und somit eine zielgruppenspezifische

[43] OECD, „Addressing the Tax Challenges of the Digitalisation of the Economy," 6.

[44] OECD, Steuerliche Herausforderungen der Digitalisierung - Zwischenbericht 2018: Inclusive Framework on BEPS, 24.

[45] Arbeitskreis Steuern der Schmalenbach-Gesellschaft für Betriebswirtschaft e.V., „Digitalisierung und Besteuerung," in Betriebswirtschaftliche Implikationen der digitalen Transformation, Hrsg. Stefan Krause und Bernhard Pellens (Wiesbaden: Springer Fachmedien Wiesbaden, 2018), 123-126.

[46] OECD, Steuerliche Herausforderungen der Digitalisierung - Zwischenbericht 2018: Inclusive Framework on BEPS, 24.

[47] Ebd., 25-26.

[48] OECD, Herausforderungen für die Besteuerung der digitalen Wirtschaft, 85.

[49] Ebd., 90.

[50] 94,5 % der App-Downloads sind Gratis-Apps

[51] OECD, Herausforderungen für die Besteuerung der digitalen Wirtschaft, 90-91.

Werbung realisieren kann. Zudem bieten viele Internetwerber die Möglichkeit, den Erfolg der Werbung zu beobachten sowie weitere Informationen herauszufinden und den Kunden bereitzustellen. Dies können beispielsweise Informationen bezüglich weiterer Interessen der Nutzer sein.[52]

Die bedeutendsten Beispiele für Online-Werbung sind die Display-Werbung sowie auch die Suchmaschinenwerbung. Während bei der Suchmaschinenwerbung Geld dafür gezahlt wird, in den Suchmaschinenergebnissen aufzutreten, wird bei der Display-Werbung das Geld dafür aufgewendet, die Werbung aufgrund eines bestimmten Nutzerverhaltens oder aufgrund bestimmter Inhalte anzuzeigen. Die Online-Werbung verzeichnet einen Boom und soll sich laut Schätzungen von 2015 auf 2021 nahezu verdoppeln, wie die untenstehende Abbildung 2 verdeutlicht. Damit wäre die Online-Werbung nach der Fernsehwerbung das zweitgrösste Werbemedium. Spitzenreiter der Online-Werbung ist die Suchmaschinenwerbung.[53]

2.2.3. Cloud Computing

Szer definiert Cloud Computing als „neuartiges Rechenparadigma, welches die Auslagerung von Daten-, Rechen- und Netzwerkressourcen mit der Möglichkeit der elastischen Ressourcenskalierbarkeit sowie der Bereitstellung dieser Ressourcen auf Abruf, mittels geringer vorausgehender Investitionskosten ermöglicht".[55] Die OECD hingegen spricht bei Cloud Computing über Online-Computerdienstleistungen[56], welche auf Abruf mithilfe der Nutzung von physischen und virtuellen Ressourcen wie Netzwerken und Server bereitgestellt werden. Ein Vorteil des Cloud Computing ist die Möglichkeit, auf Cloud-Dienste unabhängig des Standortes und mit unterschiedlichen Endgeräten zugreifen zu können. Voraussetzung dafür ist nur eine funktionierende Internetverbindung. Da der Erwerb einer eigenen IT-Infrastruktur für Unternehmen sehr kostspielig sein kann, greifen immer mehr Unternehmen zum Cloud Computing. Hierbei verteilen sich die Kosten auf viele Nutzer der Cloud und werden so geringer. Die Kostenvorteile sind insbesondere durch die Skaleneffekte bei der Errichtung sowie der Ausnutzung der Kapazität der Rechner zu erklären. Dies wird durch die Flexibilität des Cloud Computing ermöglicht, da Kunden ihre Kapazität und den Bedarf der Speicherleistung flexibel anpassen können.[57]

2.2.4. Partizipative Netzwerkplattformen

Partizipative Netzwerkplattformen stellen eine Plattform bereit. Deren User helfen dabei, einen nutzergenerierten Inhalt[58] zu entwickeln, bereitzustellen sowie auch zu kommentieren und zu verbreiten. Es gibt verschiedene Arten der Netzwerkplattformen. Während Wikis und Blogs textbasierte Plattformen sind, zählen Social-Bookmarking-Sites zu den gruppenbasierten Plattformen. Weiters gibt es Podcasts, virtuelle Welten und soziale Netzwerke, welche die bekanntesten Formen von partizipativen Netzwerkplattformen darstellen. Die Nutzer dieser Plattformen erzielen aus der Benutzung und des Kreierens der Inhalte keinen Gewinn. Die Plattformen selbst können jedoch auf unterschiedliche Weise Gewinn erwirtschaften. Dazu zählen beispielsweise der Verkauf von Nutzerdaten, Spenden, werbefinanzierte Modelle oder ein Verkauf von Waren und Dienstleistungen an die Nutzer.[59]

2.2.5. E-Commerce

Die OECD definiert E-Commerce als den Kauf oder Verkauf von Waren oder Dienstleistungen, welcher über Computernetze abgewickelt wird, die speziell für den E-Commerce Handel geschaffen wurden. Dabei müssen Kauf/Verkauf online, Bezahlung und Lieferung aber nicht zwingend online durchgeführt werden. E-Commerce kann einerseits nur als Erleichterung der Bestellung genutzt werden, während Waren und Dienstleistungen über konventionelle Kanäle abgewickelt werden. Diese Form des E-Commerce wird Offline oder indirekter E-Commerce genannt. Andererseits kann E-Commerce den gesamten Bestell- und Lieferungsprozess abwickeln, weshalb diese Art auch als Online bzw. direkter E-Commerce bezeichnet wird. Die OECD unterscheidet E-Commerce weiters in B2C (Business-to-Consumer), B2B (Business-to-Business) und C2C Modelle (Consumer-to-Consumer). Die Digitalisierung und das Internet führen zu einer signifikanten Zunahme des E-Commerce, da mithilfe des Internets die Transaktionen viel schneller und auch kostengünstiger durchgeführt werden können. Des Weiteren trägt das Internet auch dazu bei, dass Unternehmen, welche bisher nur regional operierten, ihre Geschäfte nun in andere Märkte ausweiten können und somit eine viel grössere Reichweite haben.[60]

2.2.6. Hochgeschwindigkeitshandel

Mithilfe hochmoderner Technologie wie beispielsweise Computeralgorithmen wird ein Handel von Wertpapieren mit extrem hoher Geschwindigkeit ermöglicht. Diese Algorithmen nehmen Kaufs- und Verkaufsentscheidungen vor, welche im Millionstel und Milliardstel Bereich von Sekunden liegen. Der Hochgeschwindigkeitshandel ermöglicht es, eine

[52] Ebd.

[53] Ebd., 91-92; Christian Allner, „300 Milliarden Euro: Steigender Anteil von Werbung in Sozialen Medien (inkl. Podcast)," Zugriff am 07.03.2020, https://socialmediastatistik.de/werbung-social-media/.

[54] Allner, „300 Milliarden Euro: Steigender Anteil von Werbung in Sozialen Medien (inkl. Podcast)."

[55] Benjamin Szer, Cloud Computing und Wissensmanagement: Bewertung von Wissensmanagementsystemen in der Cloud (Hamburg: Diplomica Verlag, 2013), 48.

[56] Beispielsweise Rechenleistung, Speicher, Software und Datenverwaltung

[57] OECD, Herausforderungen für die Besteuerung der digitalen Wirtschaft, 92-93.

[58] User Created Content: Text, Audio, visuelle Inhalte

[59] OECD, Herausforderungen für die Besteuerung der digitalen Wirtschaft, 96.

[60] Ebd., 85-87.

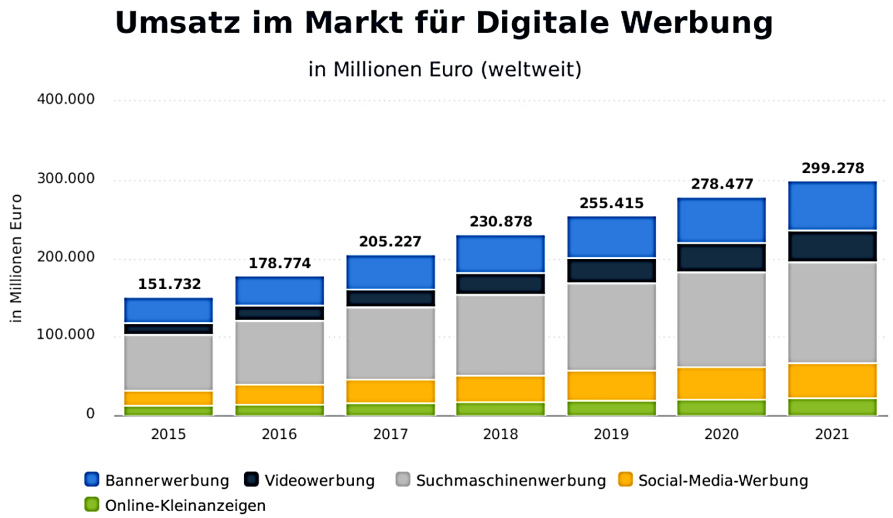


Abbildung 2: Umsatz Online-Werbung[54]

Vielzahl von Orders, welche oft kleine Stückzahlen aufweisen, an die Märkte zu schicken. Diese werden auf Hochleistungscomputern durchgeführt. Diese Hochleistungscomputer sind in der Lage, Marktdaten zu analysieren und geringe Kursschwankungen, welche oft nur Millisekunden vorhanden sind, auszunutzen. Hochfrequenzhändler zeichnen sich dadurch aus, dass sie ihre Marktpositionen nur sehr kurze Zeit halten und durch vielfache, aber kleine Transaktionen von geringen Kursänderungen profitieren können. Der Ort des Servers spielt im Hochfrequenzhandel eine bedeutende Rolle, da ein Server in der Nähe der jeweiligen Börse einen zentralen Vorteil bringen kann, um schneller als der Konkurrent zu sein.[61]

Wie die untenstehende Abbildung 3 zeigt, nimmt der Anteil des Hochfrequenzhandels deutlich zu. Insbesondere in den USA werden 51 % des Aktienhandels mittels Hochfrequenzhandel getätigt, gefolgt von Japan und Europa. Abbildung 4 zeigt die Dauer einer Transaktion im Hochfrequenzhandel, welche im Jahr 2006 von anfangs 0,021 Sekunden auf 0,00025 Sekunden derzeit sank.

2.2.7. Online-Zahlungsdienste

Online-Zahlungsdienste haben das Ziel, Finanztransaktionen zu vereinfachen, indem kein Austausch von sensibler Finanzinformation zwischen den Transaktionspartnern mehr stattfinden muss. Sie bieten eine sichere Möglichkeit, Online-Zahlungsdienste durchzuführen. Die Zahlungsanbieter fungieren dabei als Intermediär zwischen Online-Käufer und Verkäufer. Sie nehmen also Zahlungen der Käufer an, welche durch eine Vielzahl von Möglichkeiten getätigt werden können, wie beispielsweise Kreditkartenzahlungen oder auch Möglichkeiten der Bankzahlungen wie Lastschrift oder Direktüberweisung. Danach werden diese Zahlungen verarbeitet und auf das Konto des Verkäufers transferiert. Insbesondere die Sicherheit aufgrund fehlendem Austausch sensibler Information, die Schnelligkeit des Transfers im Gegensatz zu traditionellen Methoden und auch die Möglichkeit, in verschiedenen Währungen zu zahlen, bietet zahlreiche Vorteile für die Nutzer solcher Online-Zahlungsdienste. Die Geschäftsmodelle dieser Online-Zahlungsdienste können sehr unterschiedlich sein. Einige Anbieter verlangen Gebühren pro Transaktion, während andere Anbieter einen Prozentsatz des Transaktionsbetrages oder einen Pauschalbetrag im Monat verlangen.[64] Als Alternative werden von der OECD auch virtuelle Währungen genannt, welche zum Kauf für Waren und Dienstleistungen genutzt werden können und erst durch die digitale Wirtschaft entstanden sind.[65]

2.3. Probleme und Herausforderungen bei der Besteuerung digitaler Geschäftsmodelle

Während die Digitalisierung viele Vorteile mit sich bringt, stellt sie die nationale wie auch die internationale Steuerpolitik vor grosse Herausforderungen. Es stellt sich hierbei die Frage, ob das momentane internationale Steuerrecht noch in der Lage ist, die Digitalisierung und damit entstandene Geschäftsmodelle richtig zu erfassen. Zudem resultiert zunehmend die Frage, wie die Rechte der Besteuerung zwischen Quellen- und Ansässigkeitsstaat aufzuteilen sind und wie zukünftig bestimmt werden kann, wo die Wertschöpfung steuerlich stattfindet. Zudem ergänzt die OECD, dass sich zusätz-

[61] Ebd., 95.

[62] Dyrk Scherff, „Millionen in Millisekunden: Hochfrequenzhandel," Zugriff am 20.02.2020, https://www.faz.net/aktuell/finanzen/fonds-mehr/hochfrequenzhandel-millionen-in-millisekunden-12803798-p2.html.

[63] Ebd.

[64] OECD, Herausforderungen für die Besteuerung der digitalen Wirtschaft, 88-89.

[65] Ebd., 90.

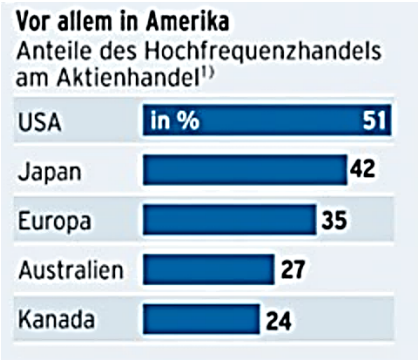


Abbildung 3: Anteil Hochfrequenzhandel[62]

Abbildung 4: Dauer Hochfrequenzhandel[63]

liche Möglichkeiten zur doppelten Nichtbesteuerung bieten und deshalb auch BEPS-Themen zu klären sind.[66]

Bereits im Endbericht des BEPS-Aktionspunktes 1 macht die OECD darauf aufmerksam, dass die Abgrenzung der digitalen Wirtschaft zur traditionellen Wirtschaft immer schwerer wird bzw. unmöglich ist. Dies vor allem aufgrund der Tatsache, dass die „normale" Wirtschaft immer mehr durch die digitale Wirtschaft abgebildet wird. Dieses so genannte „Ringfencing" ist nicht möglich, da keine klaren Grenzen herausgearbeitet werden können.[67] Die OECD weist zudem darauf hin, dass die Änderungen des Steuersystems somit auch alle bestehenden nicht digitalen Unternehmen betreffen können.[68] Auch Wünnemann macht darauf aufmerksam, dass die Digitalisierung die Art des Wirtschaftens wesentlich verändert und die digitale Wirtschaft nicht einfach von der traditionellen Wirtschaft unterschieden werden kann.[69] Versin und Hidien sehen eine Abgrenzung deshalb als sehr schwierig, da die traditionelle Wirtschaft vermehrt auf digitale Aspekte wie beispielweise Internetpages setzt.[70]

Während die Hauptintention des internationalen Steuerrechts die Vermeidung von Doppelbesteuerung ist, so verfolgt es im Zeitalter der Digitalisierung vor allem das Ziel, die Vermeidung von Missbräuchen und der doppelten Nichtbesteuerung. Dies wird bei der Betrachtung der neu entwickelten Präambeln der Doppelbesteuerungsabkommen (DBA) ersichtlich, welche oft auch schon ohne Bezug zur Vermeidung der Doppelbesteuerung implementiert werden. Im Fokus des internationalen Steuerrechts liegen vor allem die multinationalen und digitalen Konzerne. Daraus resultiert, dass die Neuheiten im internationalen Steuerrecht fast ausschliesslich Verschärfungen darstellen.[71] Faulhaber ergänzt, dass sich die Debatte zur Besteuerung der Digital Economy zu sehr auf einzelne Technologiekonzerne konzentriert, anstatt einen globalen Konsens zu den Problemen im Steuerrecht zu finden. Nicht die GAFA-Konzerne selbst stellen das internationale Steuerrecht vor Herausforderungen, sondern die hoch digitalisierten Geschäftsmodelle dieser Tech-Unternehmen.[72] Im nachfolgenden Kapitel werden die Probleme und Herausforderungen, welche durch digitalisierte Geschäftsmodelle hervorgerufen werden, näher erläutert. Dies sind die Vereinfachung von Gewinnverlagerungs- und Gewinnvermeidungsstrategien, das Konzept der physischen Präsenz, der komplexe Beitrag zur Wertschöpfung sowie die Qualifizierung der Einkünfte und Zahlungen.

2.3.1. Vereinfachung von Gewinnverlagerungs- und Gewinnvermeidungsstrategien

Seit Veröffentlichung der ersten Berichte der OECD aufgrund des BEPS-Projektes im Jahr 2014 werden die Gewinnvermeidungs- und Gewinnverlagerungsstrategien der MNE heiss diskutiert. Im Rahmen des BEPS-Projektes wird seitens der OECD ein 2000-seitiger Abschlussbericht, welcher die Gewinnverkürzungen und Gewinnverlagerungen von MNE näher beleuchtet, veröffentlicht. Nicht grundlos wurde die Digitalisierung und die digitale Wirtschaft als Aktionspunkt 1 des 15-Punkte Programmes initiiert, da die digitale Wirtschaft seit jeher im Mittelpunkt der BEPS-Arbeiten steht.[73] Die 15-Punkte des BEPS-Programmes werden in der nachfolgenden Abbildung 5 näher dargestellt.

Das Hauptproblem der Gewinnverlagerungs- und Gewinnvermeidungsstrategien ist laut Wenz die sogenannte Steuersatzarbitrage. Sie wird in Abbildung 6 näher dargestellt. Hierbei wird der Gewinn der B-GmbH durch hohe Zinsen oder Lizenzgebühren an die A-GmbH wesentlich vermindert, wobei folglich nur noch auf einen kleinen Restgewinn der Hochsteuersatz gezahlt werden muss. Der Gewinn der A-GmbH besteht somit aus den Lizenzeinkünften der B-GmbH. Diese Lizenzeinkünfte werden jedoch in Land A, als Niedrigsteuerland, gering oder gar nicht besteuert.[75]

[66] Ebd., 146-147.
[67] Ebd., 84.
[68] Ebd., 12.
[69] Monika Wünnemann, „Herausforderungen der Besteuerung der digitalen Wirtschaft im Jahr 2019," IStR, Nr. 4 (2019): 134.
[70] Versin und Hidien, „Digitales Geschäftsmodell als steuerlicher Wettbewerbsfaktor," 759.
[71] Goetz Kempelmann, „The Future of International Tax," IStR, Nr. 16 (2019): 664.
[72] Faulhaber, „Taxing Tech: The Future of Digital Taxation," 40.
[73] Georg Kofler, Gunter Mayr und Christoph Schlager, „Digitalisierung und Betriebsstättenkonzept," RdW, 5b (2017): 371- 372.
[74] Wenz, „The Future of International Taxation," 13.
[75] Martin Wenz, „Internationale Unternehmensbesteuerung: Holding- und

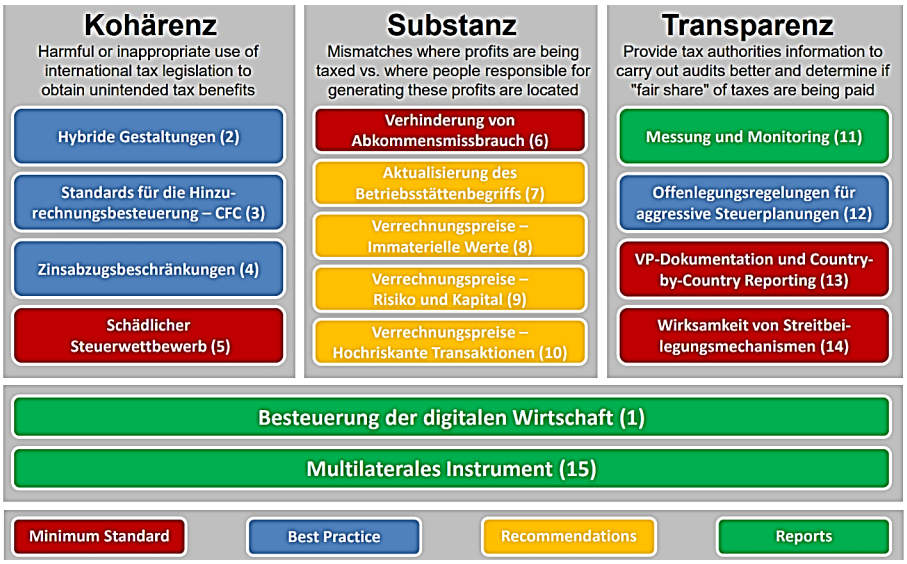


Abbildung 5: 15-Punkte des BEPS-Programmes[74]

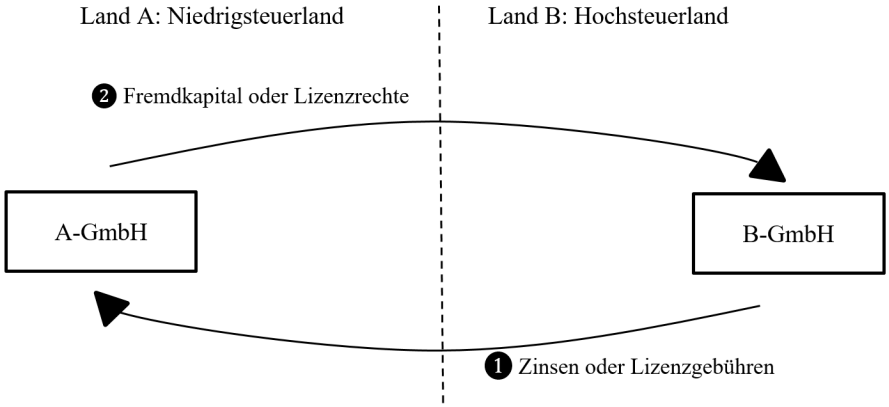


Abbildung 6: Steuersatzarbitrage
Eigene Abbildung in Anlehnung an Wenz[76]

Hinzukommend berichten Medien über die Steuervermeidungspraktiken grosser IT-Konzerne und sorgen damit für Aufruhr. Die Europäische Kommission kommt zum Entschluss, dass Apple aufgrund von Verrechnungspreisregeln und damit einhergehender Steuervergünstigungen in Irland rund 13 Mrd. Euro an Steuern gespart hat. Auch Google hat in den USA einen medialen Disput über die Steuergestaltungsmodelle der IT-Konzerne ausgelöst. Diese Diskussion schwappt nach und nach auf Europa über und beherrscht auch hier den politischen Diskurs. Frankreich kündet rund 2,5 Mrd. Euro Strafe gegenüber den GAFA-Konzernen an. Laut Kofler, Mayr und Schlager ist die Intention der Länder bezüglich solcher Vorgehensweisen verständlich, jedoch ist diese rechtsstaatlich noch völlig offen und unklar.[77]

Auch der G7-Gipfel in Chantilly befasst sich mit der Digitalisierung im Hinblick auf das internationale Steuerrecht. Hierzu wurde als Hauptproblem genannt, dass die Digitalisierung die Verlagerung von Gewinnen in Niedrigsteuerländer immens vereinfacht. Dies führt zudem zu einer Ungleichbehandlung der Unternehmen, da die erhobene Steuer bei Unternehmen mit digitalen Geschäftsmodellen im Durchschnitt viel geringer ist als bei Unternehmen mit traditionellen Geschäftsmodellen.[78] Die effektiven Steuersätze von MNE sind rund 4 bis 8,5 % niedriger als die von lokalen Unternehmen mit inländischer Geschäftstätigkeit. Digitale Top-Unternehmen haben zudem ein Umsatzwachstum von durchschnittlich 14 %, während andere MNE ein Umsatzwachstum von 0,2 % bis 3 % aufweisen. Laut Schätzungen sind Umsätze

Finanzierungsgesellschaften," University of Liechtenstein, 37.
[76] Ebd.
[77] Kofler, Mayr und Schlager, „Digitalisierung und Betriebsstättenkonzept," 369-370.
[78] G7, „Working for a fairer Capitalism: Press Kit," Chantilly 17-18 July 2019, 11.

mit rund 650 Mrd. US-Dollar von diesen BEPS-Aktionen betroffen. Diese Zahlen weisen auf die erhebliche Relevanz des BEPS-Problems hin und bestätigen die fortlaufende Steigerung der Zahlen über die letzten Jahre hinweg.[79] Wie aus der nachfolgenden Abbildung 7 aus dem Final Report des BEPS-Aktionspunktes 11 sichtbar wird, hat sich der Körperschaftsteuersatz aus dem Jahr 2000, verglichen mit dem Jahr 2014, in fast allen Ländern wesentlich verringert.

Boehle und Scholtholt beschreiben das Problem wie folgt: „Den Kern der derzeitigen Problemauffassung bildet die Annahme, dass sich internationale Unternehmen die Heterogenität der weltweiten Steuersysteme zu Nutze machen, um ihre Steuerlast zu miniminieren." Insbesondere die Digitalisierung ermöglicht laut den Autoren ein einfacheres Ausnutzen von Gewinnverlagerungs-, Gewinnverkürzungs-, und Gewinnvermeidungsstrategien, in dem eine Trennung des Ortes der eigentlichen wirtschaftlichen Tätigkeit und des Ortes der Besteuerung möglich wird. Dadurch kann die Steuerlast bedeutsam minimiert oder sogar komplett vermieden werden.[81] Die OECD hebt hervor, dass die Digitalisierung nicht komplett neue Probleme bezüglich BEPS schafft, sondern die BEPS-Probleme zusätzlich weiter verschärft.[82] Becker und Englisch machen darauf aufmerksam, dass die bisherigen BEPS-Massnahmen zu kurz greifen, um digitale Konzerne in Europa zu besteuern. Ausschlaggebend dafür ist die nur behutsame Ausweitung des Betriebsstättenbegriffs, welche trotzdem nicht in der Lage ist, die Internet-Riesen ohne physische Präsenz zu besteuern.[83]

Bendlinger stimmt zu, dass es Unternehmen der digitalen Wirtschaft um einiges leichter fällt, Gewinne dort anfallen zu lassen, wo diese den „geringstmöglichen steuerlichen Schaden" anrichten. Dass sich die Grundsätze des internationalen Steuerrechts nicht mit der digitalen Wirtschaft vereinen lassen, führt zu zahlreichen Möglichkeiten von Steueroptimierungen.[84] Ein Beispiel für ein typisches Steuersparmodell der Unternehmen ist das bekannte „Double Irish Dutch Sandwich". Ein Unternehmen, welches dieses Modell nutzt, ist beispielsweise Google. Aufgrund der Unterschiede der nationalen Gesetzgebungen ist es für Konzerne einfach, Schlupflöcher zu finden, den Gewinn zu minimieren und somit einer Besteuerung zu entkommen.[85] Im Folgenden wird nun das „Double Irish Dutch Sandwich" als Beispiel für eine Steuervermeidungsstrategie der digitalen Konzerne vereinfacht erläutert. Die folgende Abbildung 8 veranschaulicht das „Double Irish Dutch Sandwich" grafisch.

Anfangs überträgt eine Gesellschaft, beispielsweise in den USA ansässig, immaterielle Wirtschaftsgüter (wie Patente und Lizenzen) an eine irische Gesellschaft. Hierfür wurde die irische IP-Holding-Gesellschaft absichtlich mit einer Doppel-Ansässigkeit ausgestattet. Sie ist sowohl in Irland als auch auf den Bermudas ansässig. Da diese Gesellschaft einerseits in Irland registriert ist, ist sie aus Sicht der USA eine irische Gesellschaft. Andererseits befindet sich das Management, also der Ort der Geschäftsleitung, auf den Bermudas, weshalb sie nach irischem Steuerrecht eine Gesellschaft der Bermudas darstellt.[86]

Des Weiteren werden unter dieser oben genannten IP-Holding-Gesellschaft noch zwei weitere Tochtergesellschaften gegründet. Eine Tochtergesellschaft befindet sich in den Niederlanden und eine weitere in Irland.[87] Diese Tochtergesellschaft in Irland ist für das operative Geschäft zuständig und soll somit lediglich eine Ansässigkeit in Irland und der EU begründen. Sie verwaltet das immaterielle Vermögen basierend auf den Lizenzvereinbarungen und erzielt damit sehr hohe Umsätze.[88] Da diese Gesellschaft in den anderen europäischen Staaten keine weiteren Niederlassungen aufweist, besteht in diesen Ländern kein Anspruch auf Gewinnbesteuerung. Obwohl die hohen Umsätze der operativen Gesellschaft in Irland sehr wohl besteuert werden, kann die Besteuerung auf ein Kleinstes minimiert werden. Dies ist möglich, da diese Gesellschaft lediglich die Verwertung immaterieller Wirtschaftsgüter vornimmt, welche der IP-Holding-Gesellschaft zuzurechnen sind. In diesem Fall müssen hohe Lizenzzahlungen an diese IP-Holding-Gesellschaft gezahlt werden, weshalb die Steuer der operativen irischen Gesellschaft auf nahezu null minimiert wird.[89]

Laut Aigner et al. sind Lizenzzahlungen gemäss irischem Steuerrecht sehr wohl steuerbar, da sie an eine ausländische (hier Bermudas) Gesellschaft fliessen und somit der Quellenbesteuerung unterlägen. Um dies jedoch zu vermeiden, wurde die zuvor erwähnte niederländische Gesellschaft zwischengeschaltet. Diese Zwischengesellschaft hat die Lizenzen sublizensiert und verhindert so eine Besteuerung, da innerhalb Europas eine Zinsen-Lizenzgebühren-Richtlinie vorhanden ist, welche von der Quellensteuer befreit. Es wurde deshalb eine niederländische Zwischengesellschaft gewählt, da nach nationalem niederländischem Recht keine Quellensteuer auf Zahlungen an ausländische Gesellschaften anfallen. Die Zahlungen innerhalb der beiden irischen Gesellschaften hätten zu einer Quellensteuer geführt, da die IP-Holding-

[79] OECD, Measuring and Monitoring BEPS, Action 11 - 2015 Final Report: OECD/G20 Base Erosion and Profit Shifting Project (Paris: OECD Publishing, 2015), 15-16; Tørsløv, Wier und Zucman, „40% of multinational profits are shifted to tax havens each year."

[80] OECD, Measuring and Monitoring BEPS, Action 11 - 2015 Final Report, 138.

[81] Boehle und Scholtholt, „Gestaltungsoptionen zur gerechten Besteuerung von Digitalkonzernen," 920.

[82] OECD, Steuerliche Herausforderungen der Digitalisierung - Zwischenbericht 2018: Inclusive Framework on BEPS, 18.

[83] Johannes Becker und Joachim Englisch, „Ein grösseres Stück vom Kuchen: Besteuerung der Gewinne von Google und Co.," Wirtschaftsdienst 97, Nr. 11 (2017): 803.

[84] Stefan Bendlinger, „(Un-)Sinn der digitalen Betriebsstätte: Richtlinienpaket der EU zur Besteuerung der Digital Economy," SWI 28, Nr. 6 (2018): 270.

[85] Kofler, Mayr und Schlager, „Digitalisierung und Betriebsstättenkonzept," 372.

[86] Dietmar Aigner et al., „Die Veränderung des Steuerrechts durch die digitale Wirtschaft," in Digitale Transformation im Wirtschafts- & Steuerrecht, Hrsg. Elias Felten et al., 1. Aufl. (Wien: Linde, 2019), 307-308.

[87] Deshalb „Double Irish"

[88] Bereitstellung von Werbedienstleistungen beispielsweise

[89] Aigner et al, „Die Veränderung des Steuerrechts durch die digitale Wirtschaft," in Digitale Transformation im Wirtschafts- & Steuerrecht, 307-308.

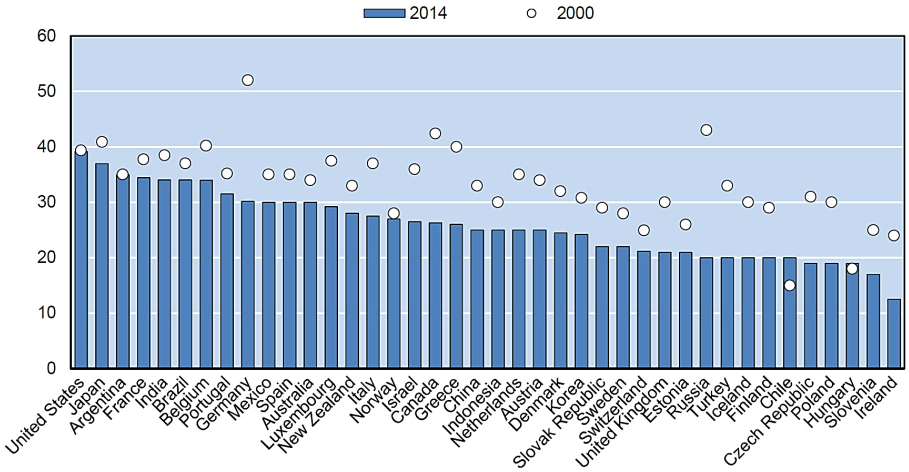


Abbildung 7: Veränderung Körperschaftsteuersätze[80]

Gesellschaft laut irischem Recht[90] auf den Bermudas ansässig ist. Durch die Zwischenschaltung[91] konnte dies vermieden werden.[92]

Abschliessend werden die Gewinne, welche noch unversteuert sind, vollkommen der IP-Holding Gesellschaft zugerechnet. Diese Gewinne bleiben jedoch steuerfrei, da die Bermuda-Inseln als Steueroase keine Körperschaftsteuer erheben und Irland die IP-Holding-Gesellschaft als nicht in Irland, sondern auf den Bermudas ansässig sieht und somit nicht besteuert.[93]

2.3.2. Konzept der physischen Präsenz

Nach momentanem Recht und laut OECD Musterabkommen kann ein Staat dann Körperschaftsteuer erheben, wenn eine physische Betriebsstätte im jeweiligen Staat vorhanden ist. Folglich kann dieser Staat auf die Gewinne dieser Betriebsstätte den Körperschaftsteuersatz des eigenen Landes erheben.[95] Für Kofler, Mayr und Schlager ist der Begriff der Betriebsstätte im internationalen Steuerrecht von signifikanter Relevanz, da dieser Nexus-indizierend agiert und somit die Schwelle einer ökonomischen Verbundenheit zweier Staaten darstellt. Diese Schwelle übertretend wird eine Besteuerung im Quellenstaat ermöglicht.[96] Entfacht wurde die steuerpolitische Diskussion rund um den Betriebsstättenbegriff aufgrund des BEPS-Aktionspunktes 1, der das Ziel verfolgt, die Steuereinnahmen der Staaten zu sichern. Bereits damals stehen dabei die Geschäftsmodelle der amerikanischen digitalen Unternehmen im Mittelpunkt, welche ohne das Aufweisen einer Betriebsstätte in einem Land eine wirtschaftliche Aktivität ermöglichen können.[97] Für Kofler, Mayr und Schlager sind die Grenzen des internationalen Steuerrechts nun sichtbar, da die digitale Wirtschaft „mit neuen Geschäftsmodellen, auf der Basis moderner Informations- und Kommunikationstechnologie mit ihrer Dezentralisierung und Entkopplung von einer physischen Präsenz“ eine grosse Herausforderung darstellt.[98]

Die Frage, inwieweit die momentane Betriebsstätten-Definition ihre zu Grunde liegenden Prinzipien noch erfüllen kann, steht im Abschlussbericht des BEPS Aktionspunktes 1 im Raum. Durch die Digitalisierung wird es Unternehmen inzwischen leicht ermöglicht, ohne eine Betriebsstätte wesentlich am Wirtschafts- und Geschäftsleben eines anderen Staates teilzunehmen. Dies wird laut OECD dadurch ermöglicht, dass Tätigkeiten im Zeitalter der Digitalisierung nun immer öfter standort-ungebunden durchgeführt werden können. Weiters können durch digitale Geschäftsmodelle Grenzen und Entfernungen überwunden werden und somit die Kunden einfach erreicht, angesprochen und erweitert werden. Während derartige Massnahmen früher noch Personal und Infrastruktur im jeweiligen Land benötigten, kann dies heutzutage grenzüberschreitend aus anderen Ländern durchgeführt werden.[99]

Zinowsky hebt hervor, dass die grösste Herausforderung des internationalen Steuerrechts die momentanen Nexus-Regeln, also die steuerlichen Anknüpfungspunkte, sind. Diese sind nicht in der Lage, die Aktivitäten digitaler Unternehmen richtig zu erfassen und somit findet die Besteuerung nicht dort statt, wo Gewinne erzielt werden und Wertschöp-

[90] Steuerlicher Anknüpfungspunkt in Irland nur aufgrund des Ortes der Geschäftsleitung

[91] Deshalb „Dutch Sandwich“

[92] Aigner et al., „Die Veränderung des Steuerrechts durch die digitale Wirtschaft,“ in Digitale Transformation im Wirtschafts- & Steuerrecht, 307-308.

[93] Ebd.

[94] Achim Pross, „Faire Besteuerung multinationaler Unternehmen: Neue Initiativen internationaler Steuerpolitik,“ OECD, 15.

[95] Pola Schneemelcher und Paul-Jasper Dittrich, „Nutzer, Daten, Netzwerke: Ein Vorschlag zur Besteuerung der Digital-Wirtschaft im europäischen Binnenmarkt,“ Jaques Delors Institut, 3.

[96] Kofler, Mayr und Schlager, „Digitalisierung und Betriebsstättenkonzept,“ 370.

[97] Wünnemann, „Herausforderungen der Besteuerung der digitalen Wirtschaft im Jahr 2019,“ 134.

[98] Kofler, Mayr und Schlager, „Digitalisierung und Betriebsstättenkonzept,“ 369.

[99] OECD, Herausforderungen für die Besteuerung der digitalen Wirtschaft, 148-149.

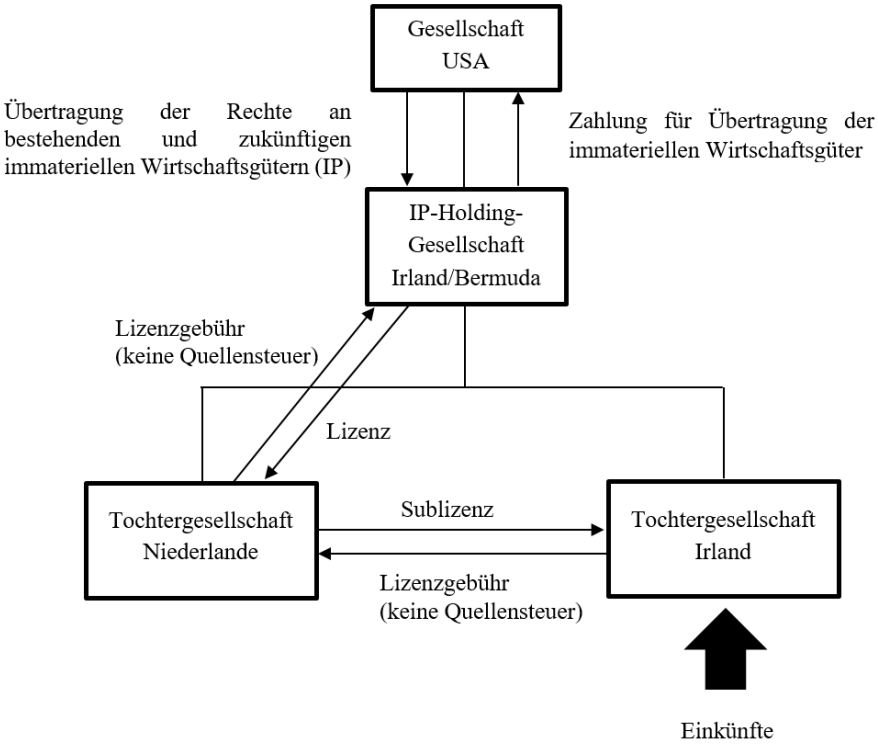


Abbildung 8: Double Irish Dutch Sandwich
Eigene Abbildung in Anlehnung an Pross[94]

fung generiert wird. Dass das internationale Steuerrecht oft als unfair und nicht sachgerecht wahrgenommen wird, liegt nicht nur an diversen Gewinnverlagerungsmassnahmen multinationaler Unternehmen, sondern auch daran, dass das Konzept der physischen Präsenz überholt ist und nicht sinngemäss an digitale Unternehmen angepasst werden kann. Zinowsky weist zudem darauf hin, dass des Öfteren auch trotz Vorliegen eines Nexus die Verteilung des Besteuerungssubstrats der Digitalunternehmen als ungerecht wahrgenommen wird.[100] Auch die OECD stimmt zu, dass im heutigen Zeitalter der Digitalisierung die bestehenden Steuerrechte nicht mehr ausschliesslich auf dem Prinzip der physischen Präsenz aufbauen können. Mit zunehmender Globalisierung ist dies nicht mehr zeitgemäss und sichert keine faire Besteuerung mehr, da diese Regeln bereits aus den 1920er Jahren stammen.[101] Die Vorschläge der OECD führen laut Bendlinger wohl dazu, dass „das bekannte und bewährte Betriebsstättenkonzept jedenfalls ausgedient hat".[102]

Das BEPS-Projekt versucht mit dem Aktionspunkt 7 den Begriff der Betriebsstätte zu erweitern und somit die Schwelle zu senken, ab derer der Quellenstaat die Besteuerung von Gewinnen durchführen kann. Jedoch trifft dieser Aktionspunkt 7 nicht direkt die Probleme der digitalen Geschäftsmodelle, welche sich eben dadurch auszeichnen, dass gar keine physische Präsenz des Unternehmens vorhanden ist.[103]

2.3.3. Komplexer Wertschöpfungsbeitrag

Schneemelcher und Dittrich machen darauf aufmerksam, dass „das Recht, Steuern auf den Gewinn eines Unternehmens zu erheben, das Land hat, in dem die Wertschöpfung stattfindet – besteuert wird dort, wo Wert geschaffen wird". Jedoch sind die internationalen Grundsätze schon lange nicht mehr übereinstimmend mit den neuen Formen der Wertschöpfung, da der immer grösser werdende Wertschöpfungsanteil immaterieller Güter zunimmt, so die Autoren.[104] Die Wertschöpfung definiert den zusätzlichen Beitrag, den Unternehmenseinheiten leisten, um innerhalb einer Produktion zur Entstehung und Qualität eines Produktes beizutragen. Unternehmenseinheiten, welche keinen wirklichen Beitrag zur Entstehung eines Produktes leisten, wird auch kein anteiliger Gewinn zugeordnet. Dies ist vor allem im BEPS-Programm von grosser Bedeutung, da Steueroasen ohne wirkliche Geschäftstätigkeit somit von der Zuordnung des Gewinnes ausgeschlossen werden können.[105] Die Debatte rund um die digitale Wirtschaft betrifft nicht mehr nur

[100] Zinowsky, „Abkehr von tradierten Verrechnungspreismethoden und Einführung einer weltweiten Mindestbesteuerung," 811-812.
[101] OECD, „Secretariat Proposal for a "Unified Approach" under Pillar One," 6.
[102] Bendlinger, „BEPS 2.0 - Besteuerung digitalisierter Geschäfte - eine (un)endliche Geschichte," 110.
[103] Kofler, Mayr und Schlager, „Digitalisierung und Betriebsstättenkonzept," 372.
[104] Schneemelcher und Dittrich, „Nutzer, Daten, Netzwerke," 1-3.
[105] Schön, „Seminar F (IFA/OECD): Internationale Besteuerung der digitalisierten Wirtschaft," 648.

die Steuervermeidungspraktiken der Konzerne, sondern die Veränderungen der Wertschöpfungsprozesse und den damit einhergehenden Problemen zur Lokalisierung und Quantifizierung der Wertschöpfung, insbesondere auch innerhalb von Konzernen.[106]

Auch die OECD erwähnt 2018 in ihrem Interim Report, dass sich die Wertschöpfung stark verändert hat. Dabei erwähnt sie die Wertekette nach Porter, welche im Zeitalter der Digitalisierung nicht mehr verwendet werden kann bzw. einige grosse Probleme bereitet. Dies resultiert einerseits aus der Tatsache, dass Porters Wertekette den Informationsfluss und die Informationsverarbeitung als rein unterstützende Aktivitäten berücksichtigt, während bei digitalen Geschäftsmodellen diese Informationen immer mehr in den Vordergrund rücken, wenn nicht sogar eine Hauptrolle spielen. Weiters wurde Porters Wertschöpfungskette auf Güter, anstelle von Dienstleistungen ausgerichtet. Das dritte Problem, welches die Anpassung des Wertekettemodells an digitale Geschäftsmodelle erschwert, ist die geografische Begrenztheit.[107]

Die unter Abschnitt 2.1 erwähnten Eigenschaften von digitalen Geschäftsmodellen erhalten globale Zustimmung, jedoch herrscht weiter Diskussion darüber, inwieweit diese Eigenschaften für den Ort der Wertschöpfung bedeutend sind. Insbesondere die Eigenschaft Daten und Nutzerbeteiligung (siehe 2.1.3) entfacht Diskussionen, da kein internationaler Konsens darüber gefunden werden kann, ob und wie diese Eigenschaft einen Beitrag zur Wertschöpfung leisten kann.[108] Postler identifiziert die fehlende Erkenntnis, wie innerhalb digitaler Geschäftsmodelle Wertschöpfung stattfindet, als Hauptproblem der digitalen Zeit. Jedoch gibt es internationale Zustimmung darüber, dass im Zeitalter der Digitalisierung der Ort und Umfang der Besteuerung mit der Wertschöpfung vereint werden muss. Um Gewinne grenzüberschreitend anhand ihrer Wertschöpfung richtig zu verteilen, benötigt man ein ausgeprägtes Wissen über die digitalen Geschäftsmodelle und ihre Wertschöpfung.[109] Die Tatsache, dass die Wertschöpfung digitaler Unternehmen heutzutage vor allem auf immateriellen Wirtschaftsgütern basiert, wirft immer mehr Fragen dazu auf, wie die dazugehörigen Einnahmen zuzuordnen sind. Zudem wird die Verschiebung in Steueroasen mit immateriellen Wirtschaftsgütern noch einfacher.[110]

Einerseits vertreten einige Staaten die Ansicht, dass vor allem die Daten und Nutzerbeteiligung von wesentlicher Bedeutung für digitale Unternehmen sind und den wichtigsten Treiber der Wertschöpfung darstellen. Unternehmen haben durch die aktive Nutzerbeteiligung die Möglichkeit, immense Mengen an relevanten Daten zu sammeln und durch Netzwerkeffekte einfach weitere Nutzer zu gewinnen. Die Beteiligung der Nutzer ist für diese digitalen Geschäftsmodelle von enormer Bedeutung, da diese beispielsweise zu einer Erhöhung des Markenwerts, des Vertrauens, des Ansehens sowie zur Generierung von Netzwerkeffekten führen können. Digitale Unternehmen haben die Möglichkeit, benutzerbezogene Daten auf einer digitalen Plattform in einem Staat zu sammeln, diese dann zu analysieren und dazu zu nutzen, um direkt an die Nutzer zugeschnittene Werbung zu schalten. Diese Aktivitäten sind für einige Staaten ein wesentlicher Beitrag zur Wertschöpfung, da die Nutzer indirekt für das Unternehmen zusätzlichen Wert schaffen können und den Nutzen sowie den Wert einer Plattform enorm erhöhen können. Für diese Staaten ist „der Beitrag der Nutzer etwas wirklich Neuartiges, das über die reine Inanspruchnahme einer Dienstleistung (d. h. Zugang zu dem Geschäftsmodell) hinausgeht".[111]

Andererseits argumentieren andere Staaten, dass die Bereitstellung von Daten und die Nutzerbeteiligung eine Transaktion zwischen Geschäftspartnern darstellt und die Nutzer finanziell oder nichtfinanziell dafür entschädigt werden. Obwohl die Transaktionen zwischen dem Nutzer und dem digitalen Unternehmen sehr wohl eine Geschäftsaktivität, welche ertragsteuerlich erfassbar sein könnte, darstellt, weisen die Staaten darauf hin, dass bisherige nicht finanzielle Transaktionen heutzutage selten ertragssteuerlich erfasst werden. Nach dieser Auffassung kann die Bereitstellung von Nutzerdaten mit der Bereitstellung von Informationen von fremden Dritten in der Lieferkette verglichen werden.[112]

Daten können laut OECD einerseits eine Vorleistung im Wertschöpfungsprozess darstellen, andererseits können Daten aber sehr wohl auf unterschiedlichste Art einen wesentlichen Beitrag zur Wertschöpfung leisten. Die zunehmende Relevanz der Daten innerhalb des Wertschöpfungsprozesses stellt die Gesetzgeber vor neue Herausforderungen. Für die OECD bleibt zu klären, inwieweit Daten in einem Land, in denen sie erhoben werden, besteuert werden sollen bzw. wie Daten zukünftig steuerrechtlich qualifiziert und bewertet werden können.[113] Für Junkers werden digitale Geschäftsmodelle nicht nur durch die Erhebung und Verwertung von Nutzerdaten erfolgreich, sondern vor allem durch die Entwicklung eines aussergewöhnlichen und bedeutenden Geschäftsmodelles für Kunden. Nur durch ein gut durchdachtes Geschäftsmodell lassen sich bedeutende Mengen an Daten sammeln sowie eine erfolgreiche Behauptung am Markt möglich machen. Der bedeutendste Teil einer nutzerbasierten Wertschöpfung liegt also in der Erfindung und Entwicklung des digitalen Geschäftsmodelles, an welchem andere Staaten partizipieren wollen, ohne wirklich ausschlaggebend für den Erfolg zu sein.[114] Auch Boehle und Scholtholt stimmen zu,

[106] Arbeitskreis Steuern der Schmalenbach-Gesellschaft für Betriebswirtschaft e.V., „Digitalisierung und Besteuerung," in Betriebswirtschaftliche Implikationen der digitalen Transformation, 125.

[107] OECD, Steuerliche Herausforderungen der Digitalisierung - Zwischenbericht 2018: Inclusive Framework on BEPS, 35-36.

[108] Ebd., 25.

[109] Steffen Postler, „Die Digitalisierung traditioneller Geschäftsmodelle: Ein Ansatz zum Umgang mit steuerlichen Verrechnungspreisen," IStR, Nr. 3 (2019): 103.

[110] Versin und Hidien, „Digitales Geschäftsmodell als steuerlicher Wettbewerbsfaktor," 768.

[111] OECD, Steuerliche Herausforderungen der Digitalisierung - Zwischenbericht 2018: Inclusive Framework on BEPS, 25.

[112] Ebd.

[113] OECD, Herausforderungen für die Besteuerung der digitalen Wirtschaft, 152.

[114] Lars Junkers, „Lösungen der OECD/G20 und Europäischen Kommission

dass der Nutzerbeitrag nicht als Beitrag zur Wertschöpfung angesehen werden kann und deshalb noch lange nicht die Besteuerung im Marktstaat rechtfertigt. Auch für die Autoren besteht die Wertschöpfung nicht in der Datensammlung, sondern in der Weiterverarbeitung der Daten, welche jedoch meist am Sitz des Unternehmens durchgeführt wird. Aus diesem Grund sei eine Anpassung des Steuerrechts an digitale Geschäftsmodelle oder die Einführung einer Digitalsteuer ungerechtfertigt.[115]

Der Ausgang dieser Diskussion hat laut OECD wesentlich Einfluss darauf, ob grundsätzliche Probleme innerhalb des internationalen Steuerrechts bestehen oder die Grenzen des internationalen Steuerrechts nur aufgrund veränderter Geschäftsmodelle und digitaler Unternehmen sichtbar werden.[116]

2.3.4. Qualifizierung der Einkünfte und Zahlungen

Ein weiteres Problem der digitalen Geschäftsmodelle ist die Qualifizierung der Einkünfte und Zahlungen. Bezüglich neuer digitaler Produkte und Dienstleistungen ist unklar, wie Zahlungen richtig qualifiziert werden können. Diese Frage ist vor allem beim Geschäftsmodell des Cloud Computing von grosser Bedeutung.[117] Die OECD macht darauf aufmerksam, dass im digitalen Zeitalter neue Produkte und Dienstleistungen sowie auch Monetarisierungsmöglichkeiten entstehen. Hierbei ist es jedoch nicht klar, wie der Wert, der durch Daten beispielsweise generiert wird, qualifiziert werden kann. Die digitalen Geschäftsmodelle werfen jedoch einige Fragen, insbesondere nach innerstaatlichem Recht und nach Steuerabkommen, auf, wie beispielsweise das Einordnen in die derzeitigen Einnahmekategorien. Bei den Zahlungen in Verbindung mit Cloud Computing ist nicht klar, ob diese Zahlungen als das Mieten eines Teiles eines Servers (folglich Einstufung als Lizenzgebühren[118]), als eine Dienstleistung (folglich Einstufung als Unternehmensgewinne) oder als Gebühren für technische Dienste einzustufen sind.[119]

Während in den meisten Steuerabkommen Unternehmensgewinne nur mit Aufweisen einer Betriebsstätte in einem Staat steuerpflichtig sind, können hingegen andere Arten von Einkünften, wie beispielsweise Lizenzgebühren, sehr wohl einer Quellensteuer unterliegen. Daher kann es durchaus relevant sein, wie Zahlungen eingestuft werden. Eine Einstufung als Unternehmensgewinne oder eine andere Art von Einkünften kann innerhalb von Steuerabkommen grosse Auswirkungen haben. Aus diesem Grund muss geklärt werden, ob die bestehenden Regeln im Zusammenhang mit einigen neuen Geschäftsmodellen noch anwendbar sind.[120]

2.4. Zwischenfazit

Bauer et al. üben Kritik, dass die bisherigen Vorschläge zur Besteuerung der digitalen Welt der OECD zu umfangreich sind und meist digitale Geschäftsmodelle, welche überhaupt keine steuerrechtlichen Probleme bereiten, erfassen. Aufgrund der aufstrebenden Debatte rund um eine Besteuerung am Ort der Wertschöpfung analysieren Bauer et al. Bedingungen, die dazu führen, dass der Ort der Wertschöpfung und der Ort der Besteuerung digitaler Geschäftsmodelle auseinanderfallen. Es entstehen laut den Autoren also nur steuerrechtliche Herausforderungen hinsichtlich digitaler Geschäftsmodelle, wenn diese zwei Orte auseinanderfallen.[121] Mayr und Schlager weisen darauf hin, dass zu oft nur pauschal über alle digitalisierten Geschäftsmodelle diskutiert wird, jedoch diese sehr oft auf höchst unterschiedlichen Ansätzen beruhen und deshalb nicht pauschaliert werden können. Laut den Autoren wurde von der OECD im Endbericht des Aktionspunktes 1 zwar versucht, digitale Geschäftsmodelle abstrakt darzustellen und steuerrechtlich auf Gewinnvermeidungsstrategien zu analysieren, jedoch machen die von der OECD getroffenen Einteilungen nur wenig Sinn, falls es darum geht, zu erkennen, welche Teile dieser Geschäftsmodelle nicht sinnvoll ertragssteuerlich erfasst werden können.[122]

Laut Bauer et al. gibt es drei kumulative Bedingungen, welche zu einem Auseinanderfallen des Ortes der Besteuerung und der Wertschöpfung führen. Einerseits führt ein Unternehmen als wirtschaftliche Tätigkeit eine Datenerhebung durch. Weiters muss diese Tätigkeit in einem Land praktiziert werden, in dem das Unternehmen keinen Nexus aufweist. Zudem muss das Unternehmen diese erworbenen Daten dann indirekt oder direkt monetarisieren. Die Monetarisierung soll darauf hinweisen, dass diese Daten zur Wertschöpfung beitragen, da sie für das Unternehmen einen Wert darstellen, ohne welchen das Produkt oder die Dienstleistung nicht existieren würde. Eine reine Datenerhebung, welche nicht monetarisiert werden kann, soll laut Bauer et al. auch nicht als Wertschöpfung zählen und somit von der Besteuerung auszunehmen sein.[123] Die nachfolgende Abbildung 9 spiegelt ein prototypisches digitales Geschäftsmodell wider, welches zu steuerlichen Herausforderungen führt.

In diesem Beispiel praktiziert die A-GmbH, ansässig in Land A, eine wirtschaftliche Tätigkeit, also eine Datenerhebung (1), in Land B. Weder die A-GmbH noch der Transakti-

zur Besteuerung der Digital Economy - Eine kritische Auseinandersetzung," ISR, Nr. 10 (2018): 372-373.

[115] Boehle und Scholtholt, „Gestaltungsoptionen zur gerechten Besteuerung von Digitalkonzernen," 923.

[116] OECD, Steuerliche Herausforderungen der Digitalisierung - Zwischenbericht 2018: Inclusive Framework on BEPS, 26.

[117] Versin und Hidien, „Digitales Geschäftsmodell als steuerlicher Wettbewerbsfaktor," 768.

[118] Innerhalb Steuerabkommen sind Lizenzgebühren definiert als die Vermietung kaufmännischer, gewerblicher oder wissenschaftlicher Ausrüstung

[119] OECD, Herausforderungen für die Besteuerung der digitalen Wirtschaft, 145-155.

[120] Ebd., 156.

[121] Georg Bauer et al., „Die Besteuerung digitaler Wertschöpfung - "Doing nothing is not an option"," DStR, 16-17 (2019): 887.

[122] Gunter Mayr und Christoph Schlager, „Digitale Geschäftsmodelle und Betriebsstättenkonzept," in Digitalisierung im Konzernsteuerrecht: Schriftenreihe zum Konzern- und Unternehmenssteuerrecht, Hrsg. Kirchmayr et al. (Linde, 2018), 22-26.

[123] Bauer et al., „Die Besteuerung digitaler Wertschöpfung - "Doing nothing is not an option"," 889-890.

[124] Ebd., 890.

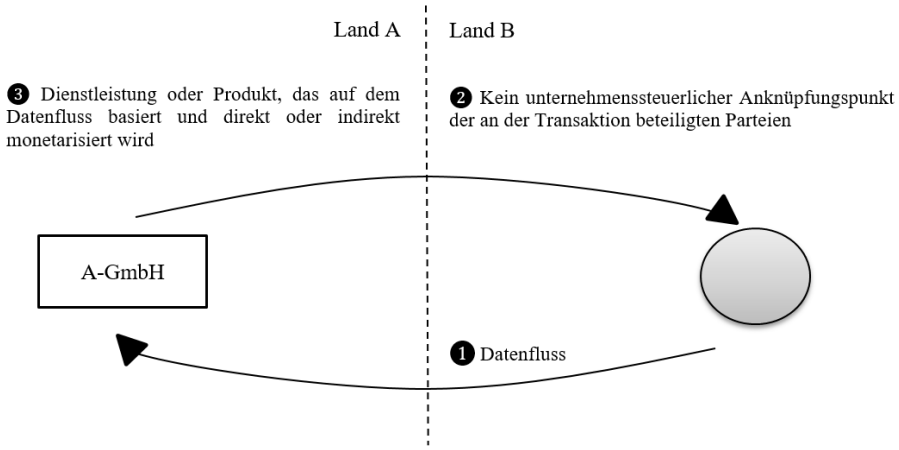


Abbildung 9: Digitales Geschäftsmodell mit steuerlichen Herausforderungen
Eigene Abbildung in Anlehnung an Bauer et. al[124]

onspartner dürfen in Land B einen Nexus begründen (2). Diese zweite Bedingung wird meist erfüllt, da der Transaktionspartner der A-GmbH ein Konsument ist. Schliesslich müssen die erhobenen Daten wesentlich zur Wertschöpfung der A-GmbH und deren Geschäftsmodell beitragen und somit eine wesentliche Verbindung zur Monetarisierung des Geschäftsmodelles abbilden (3).

Als Beispiel nennen die Autoren eine digitale Plattform, wie beispielsweise ein soziales Netzwerk, welches von der A-GmbH den Nutzern in Staat B zur Verfügung gestellt wird. Durch diesen Geschäftsfall begründen weder die A-GmbH noch die Nutzer des sozialen Netzwerks einen Nexus im Staat B. Das Geschäftsmodell der A-GmbH besteht darin, dass sie die Nutzerdaten in Staat B sammelt und somit anderen Unternehmen auf Basis der gewonnenen Daten eine individualisierte Werbeplattform zur Bewerbung ihrer Produkte zur Verfügung stellt. Dieses digitale Geschäftsmodell erfüllt die oben genannten drei Bedingungen kumulativ.[125]

Des Weiteren betrachten die Autoren in ihrer Analyse vier weitere prototypische digitale Geschäftsmodelle, welche jedoch zu keinen steuerlichen Herausforderungen führen. Diese werden in nachstehender Abbildung 10 kurz näher erläutert. Bei all diesen digitalen Geschäftsmodellen ist eine der drei Voraussetzungen nicht erfüllt, weshalb auch keine steuerlichen Herausforderungen entstehen.[126]

3. Analyse der Säule 1: Outline of the Architecture of a Unified Approach on Pillar One

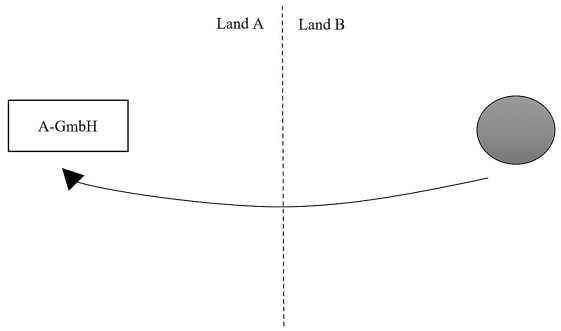
Der Inclusive Framework on BEPS der OECD, welcher sich den Problemen und Herausforderungen der digitalen Wirtschaft stellen soll, besteht aus 137 Ländern.[128] Bis Ende des Jahres 2020 soll eine langfristige Lösung herausgearbeitet werden. Säule 1 revolutioniert hierbei das internationale Steuerrecht.[129] Innerhalb der Säule 1 wird diskutiert, wie neue steuerliche Anknüpfungspunkte und damit einhergehende Gewinnverteilungsregeln geschaffen werden können. Darüber hinaus soll den Absatzmärkten ein grösserer Teil der Gewinne zugewiesen werden.[130] Für die OECD hat diese Säule 1 wesentliche Auswirkungen auf unterschiedlichste Bereiche des Steuerrechts. Dies sind beispielsweise die Zuteilung von Besteuerungsrechten zwischen Jurisdiktionen, grundlegende Konzepte des internationalen Steuerrechts wie das Konzept der physischen Präsenz und Verrechnungspreisgrundsätze sowie auch die Vermeidung unilateraler Massnahmen. Die OECD erwähnt zudem, dass auch der extreme politische Druck, digitale Geschäftsmodelle zu besteuern, immer grösser wird.[131]

Das im Jänner 2020 veröffentlichte Dokument der OECD Outline of the Architecture of a Unified Approach on Pillar One bildet die Grundlage der nachfolgenden Analyse. Die OECD macht darauf aufmerksam, dass sich die Mitglieder des Inclusive Framework on BEPS auf diesen oben erwähnten Entwurf geeinigt haben und dieser als Grundlage für weitere Verhandlungen gesehen wird, sodass im Juli 2020 eine ausgearbeitete Lösung präsentiert werden soll.[132] Im folgenden Abschnitt werden zuerst die Ziele der Säule 1 näher erläutert sowie anschliessend eine kritische Analyse der Säule 1 durchgeführt.

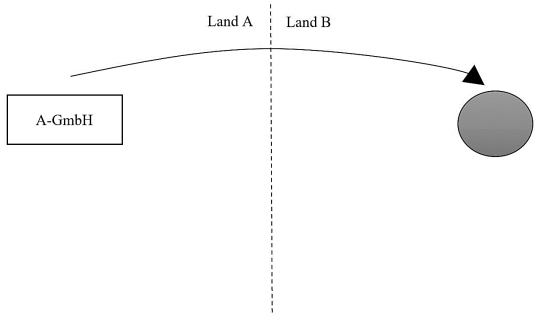
3.1. Ziele der Säule 1

Im folgenden Abschnitt werden nun die Modernisierung des internationalen Steuerrechts als auch die Vermeidung nationaler Alleingänge, als Ziele der Säule 1, näher beschrieben.

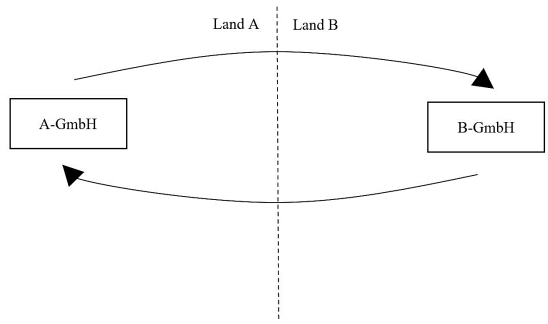
[125] Ebd.
[126] Ebd., 890-892.
[127] Ebd., 890-892.
[128] Stand Dezember 2019
[129] G7, „Working for a fairer Capitalism: Press Kit," 12.
[130] Christian Ehlermann, „Welche Richtung nimmt die Besteuerung digitaler Geschäftsmodelle?," DB, Nr. 15 (2019): M4.
[131] OECD, „Outline of the Architecture of a Unified Approach on Pillar One," 7.
[132] Ebd., 8.



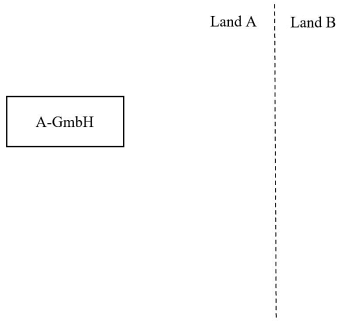
- Datensammlung in Land B → Datenfluss von Land B nach Land A
- In Land B → kein Nexus durch A-GmbH oder Transaktionspartner
- Datenerhebung ohne konkretes Ziel → keine Monetarisierung
- Kein Produkt oder DL auf Basis der gesammelten Daten



- Bereitstellung eines Produktes oder DL durch A-GmbH in Land B
- In Land B → kein Nexus durch A-GmbH oder Transaktionspartner
- Keine Datenerhebung → nur Konsum
- Keine Wertschöpfung in Land B



- Nur für B2B Konstellationen
- Bereitstellung eines Produktes oder DL durch A-GmbH in Land B
- Datensammlung in Land B → Datenfluss von Land B nach Land A
- A-GmbH → Nexus in Land A
- B-GmbH → Nexus in Land B



- Kein Auslandsbezug
- Digitale Technologien → nur für Optimierung unternehmensinterner Prozesse

Abbildung 10: Digitale Geschäftsmodelle ohne steuerliche Herausforderungen
Eigene Abbildung in Anlehnung an Bauer et. al[127]

3.1.1. Modernisierung des internationalen Steuerrechts

Die Modernisierung des internationalen Steuerrechts schreitet in den letzten zehn Jahren in grossen Schritten voran. Wesentliche Veränderungen wie die Abschaffung des Bankgeheimnisses, eine Verbesserung der Zusammenarbeit im Steuerbereich sowie auch die Einführung des BEPS-Projektes wurden umgesetzt. Um das internationale Steu-

errecht zu stabilisieren und die Effektivität zu gewährleisten, sind noch weitere Vorschriften nötig. Durch die Digitalisierung sind wesentliche Grundkonzepte des internationalen Steuerrechts gefährdet, weshalb dringender Lösungsbedarf herrscht. Aufgrund der zunehmenden nationalen Gesetze zur Besteuerung der digitalen Wirtschaft ist es von wesentlicher Bedeutung, dass Unternehmen auch in Staaten, in denen Benutzer ansässig sind, aber keine physische Präsenz besteht, besteuert werden können. Dieses Ziel soll mit den Prämissen erfüllt werden, so dass Doppelbesteuerung vermieden und Steuersicherheit gegeben ist. Die Lösung soll sowohl nachhaltig sein als auch Konflikte lösen, die weiter als die Digitalisierung reichen.[133] Auch Pistone et al. erwähnen, dass das grundlegende Ziel des Unified Approach eine Besteuerung, auch ohne das Aufweisen einer physischen Präsenz, möglich machen soll.[134]

Versin und Hidien weisen darauf hin, dass die Säule 1 eine faire und effiziente Besteuerung garantieren und zudem „die wettbewerbspolitische Funktionsfähigkeit des Marktes“ nicht einschränken soll. Die Mitgliedstaaten sind sich einig, dass eine Lösung nur global erreicht werden kann, indem ein konsistentes internationales Regelwerk beibehalten wird. Zudem ist ein weiteres Ziel, die globale Wohlfahrt, die Investitionen und das Wachstum zu fördern. Es ist von grosser Relevanz, ein multilaterales Abkommen zu erreichen, da somit die Risiken der Doppelbesteuerung und die Komplexität reduziert werden können, um schliesslich ein „gerechteres, effizienteres und einfacheres Steuersystem zu schaffen“.[135]

Ehlermann erwähnt, dass trotz vieler geplanter oder bereits in Umsetzung befindlicher BEPS-Modernisierungsmassnahmen, damit gemeint sind beispielsweise Nexus-Regeln für Intellectual Property (IP)-Boxen, Transparenz durch neue Dokumentationsregeln und überarbeitete Verrechnungspreisprinzipien, offensichtlich ist, dass diese nicht ausreichend sind, um Steuervermeidungen der digitalen Konzerne einzuschränken.[136] Auch die OECD weist darauf hin, dass mehrere Vorschläge zur Besteuerung der digitalen Welt gemacht wurden, diese jedoch meistens nicht dazu in der Lage waren, hoch digitalisierte Tech-Unternehmen richtig zu erfassen. Genau aus diesem Grund initiierte die OECD den Unified Approach, um für die Zukunft gewappnet zu sein und in Zeiten der Digitalisierung ein faires und effektives Steuerrecht zu garantieren.[137]

3.1.2. Vermeidung nationaler Alleingänge

Kempelmann weist darauf hin, dass insbesondere nationale Alleingänge die Thematik der Besteuerung der digitalen Geschäftsmodelle verschärfen, da aufgrund dieser Alleingänge die OECD auf eine Beschleunigung des Prozesses pocht. Jedoch kann noch nicht abgesehen werden, wie sich die Massnahmen der OECD auf diese unilateralen Massnahmen auswirken werden und ob diese Einfluss auf das Projekt der OECD haben werden.[138] Laut Bendlinger ist es von grosser Bedeutung, dass bis Ende 2020 rascher Konsens erfolgt. Ansonsten resultieren noch weitere nationale Alleingänge einzelner Mitgliedstaaten. Diese würden zu mehr Doppelbesteuerung führen. Aufgrund der Meinungsverschiedenheiten sowie zunehmendem politischen Druck werden immer mehr nationale Alleingänge folgen.[139] Diese nationalen Alleingänge einzelner Staaten zur Besteuerung der Digital Economy führen nun zu einer Schieflage des internationalen Steuerrechts. Diese Schieflage sollte durch das BEPS-Projekt verhindert werden.[140]

Laut Faulhaber ist der OECD bewusst, falls es nicht möglich ist, einen einheitlichen Konsens zu bilden, werden Länder zunehmend mit unilateralen Massnahmen reagieren. Dies würde einerseits zur Doppelbesteuerung und zu steigenden Compliance-Kosten, andererseits zu einem Wettbewerb zwischen den Jurisdiktionen bezüglich Besteuerungsrechten führen. Das Zustandekommen dieses Projektes hat eine grosse Bedeutung für das internationale Steuerrecht, da die Länder ansonsten zahlreiche Steuern auf Technologieunternehmen erheben werden, anstatt eine effektive und globale Lösung herauszuarbeiten und umzusetzen.[141] Die OECD betont, dass zu viele unilaterale Massnahmen die Effektivität und die Nachhaltigkeit des internationalen Steuersystems gefährden sowie dem weltweiten Wachstum und den Investitionen schaden würden.[142] Die OECD stellt zudem fest, dass als Folge einer konsensbasierten Lösung dieses Projektes sowohl eine Verpflichtung zur Umsetzung des Abkommens seitens der Mitgliedstaaten besteht, als auch jegliche unilaterale Massnahmen zurückgenommen werden müssen.[143]

Laut Bendlinger ist der Inclusive Framework on BEPS am besten dazu geeignet, einen globalen Konsens zu finden.[144] Er ergänzt zudem, dass zur Vermeidung des Verteilungswettbewerbes zwischen den Staaten langfristig nur auf globaler Ebene die Besteuerung digitaler Geschäftsmodelle angegangen werden kann und die OECD das beste Gremium dafür ist.[145] Auch für Kempelmann ist der wichtigste Akteur die OECD, die sich zwar nur als „Ohr der Mitgliedstaaten“ sieht,

[133] OECD, „OECD Secretary-General Tax Report to G20 Finance Ministers and Central Bank Governors,“ October 2019, 5.

[134] Pasquale Pistone et al., „The OECD Public Consultation Document “Secretariat Proposal for a “Unified Approach” under Pillar One”: An Assessment,“ BIT 74, Nr. 1 (2020): 15.

[135] Versin und Hidien, „Digitales Geschäftsmodell als steuerlicher Wettbewerbsfaktor,“ 769.

[136] Ehlermann, „Welche Richtung nimmt die Besteuerung digitaler Geschäftsmodelle?,“ M4.

[137] OECD, „Secretariat Proposal for a “Unified Approach” under Pillar One,“ 6.

[138] Kempelmann, „The Future of International Tax,“ 663.

[139] Bendlinger, „BEPS 2.0 - Besteuerung digitalisierter Geschäfte - eine (un)endliche Geschichte,“ 110.

[140] Stefan Bendlinger, „Steht das Betriebsstätten-Konzept vor dem Aus? Möglichkeiten und Grenzen der Besteuerung der “Digital Economy”,“ WT Fachjournal, Nr. 3 (2017): 180.

[141] Faulhaber, „Taxing Tech: The Future of Digital Taxation,“ 51-52.

[142] OECD, „Secretariat Proposal for a “Unified Approach” under Pillar One,“ 4.

[143] OECD, „Outline of the Architecture of a Unified Approach on Pillar One,“ 8.

[144] Bendlinger, „BEPS 2.0 - Besteuerung digitalisierter Geschäfte - eine (un)endliche Geschichte,“ 108.

[145] Bendlinger, „(Un-)Sinn der digitalen Betriebsstätte: Richtlinienpaket der EU zur Besteuerung der Digital Economy,“ 277.

jedoch Ideen sammelt, Vorschläge unterbreitet und umsetzt. Für ihn sind die globalen Steuergesetze alternativlos und stellen einen grossen Vorteil für das internationale Steuerrecht dar. Die Abstimmung der Steuersysteme bei grenzüberschreitenden Aktivitäten ist von grosser Bedeutung und die darin involvierten Personen der OECD verfügen über ein breites Know-how.[146]

EU-Digitalsteuer

Nicht nur die OECD, sondern auch die EU versucht passende Lösungen bezüglich der Besteuerung der digitalen Unternehmen zu finden. Vorschläge zur Digitalsteuer oder zur Besteuerung anhand einer signifikanten digitalen Präsenz scheitern jedoch aufgrund politischer Unstimmigkeiten.[147] Der EU-Richtlinienvorschlag vom 21.03.2018 zur Einführung einer EU-weiten Digitalsteuer kann keine Einstimmigkeit erlangen. Lediglich Unternehmen, welche einen weltweiten Umsatz von 750 Mio. Euro aufweisen sowie innerhalb der EU digitalsteuerpflichtige Erträge von 50 Mio. Euro erwirtschaften würden, wären von dieser Digitalsteuer betroffen gewesen. Auch der Versuch einer Einführung einer Digitalwerbesteuer am 01.03.2019 konnte nicht umgesetzt werden, da keine Einigung der Finanzminister auf EU-Ebene erreicht werden konnte.[148]

Das Scheitern der Vorschläge kann auch darauf zurückgeführt werden, dass die Mehrheit der Mitgliedstaaten eine einheitliche globale Lösung auf Ebene der OECD bevorzugen würden.[149] Der Bericht des ECOFIN[150] zur Ablehnung der EU-Digitalsteuer beinhaltet neben dem Hauptgrund, dass dieser Ansatz nur auf EU-Ebene und nicht global durchgeführt wird, auch die Problematik rund um Doppelbesteuerung, zunehmender Komplexität und Rechtsunsicherheit. Mahr merkt an, dass eine Lösung auf EU-Ebene höchst wahrscheinlich in weite Ferne gerückt ist.[151] Grassl und Koch weisen darauf hin, dass einige europäische Staaten die Entwicklungen auf globaler Ebene, also die Arbeiten der OECD, verfolgen und abwarten. Falls sich hier jedoch kein einheitlicher und schneller Konsens bilden kann, hat der ECOFIN angedeutet, den Richtlinienvorschlag der EU nochmals zu bearbeiten und zu diskutieren.[152] Für Zinowsky bleibt es spannend, wie sich die Arbeiten der OECD weiterentwickeln werden. Obwohl beim G20-Finanzminister-Treffen am 09.06.2019 eine gemeinsame Abschlusserklärung erstellt wurde, bleibt offen, wie die Verlierer dieses Ansatzes der OECD, nämlich USA und China, reagieren werden. Zuvor hatten die USA wegen der geplanten Einführung einer EU-Digitalsteuer angedroht, eine Beschwerde bei der WTO (World Trade Organization) einzulegen.[153]

Für Boehle und Scholtholt zeigen Fälle wie der Disput um Steuernachzahlungen von Apple in Irland, bei dem sowohl die USA als auch der EU-Mitgliedstaat Irland gegen die Beihilfeentscheidung der Europäischen Kommission vorgehen, dass auch innerhalb der EU Steuerwettbewerb existiert. Dieser Wettbewerb und auch die Spaltung der EU-Mitglieder auf EU-Ebene tragen dazu bei, dass keine einheitliche Lösung realisierbar ist. Aus diesem Grund fokussieren sich nun viele Staaten der EU auf das Projekt der OECD, um die Unfähigkeit der EU umgehen zu können.[154] Grassl und Koch heben jedoch hervor, dass aufgrund des Scheiterns der EU-Initiative zur Besteuerung der Digitalwirtschaft bereits zehn europäische Länder eine nationale Lösung erarbeiten. Sie machen zudem darauf aufmerksam, dass obwohl manche Länder keine eigene Digitalsteuer beabsichtigen, diese Länder sehr wohl durch die unilateralen Massnahmen der anderen Länder negativ betroffen sind. Im Folgenden werden nun die unilateralen Gesetzesinitiativen von Frankreich und Österreich näher erläutert, da die Prozesse dieser Länder am fortgeschrittensten sind.[155]

Nationaler Alleingang: Frankreich

Da Frankreich innerhalb der EU seit zwei Jahren eine Digitalsteuer fordert, die Vorschläge im EU-Rat aber zurückgewiesen worden sind, reagierte Frankreich nun mit einer unilateralen Massnahme. Für Frankreichs Finanzminister Bruno Le Maire ist eindeutig, dass die derzeitige Lage, dass Digitalkonzerne zu wenig Steuern in Frankreich zahlen, nicht länger zu akzeptieren ist. Der französische Entwurf einer Digitalsteuer vom 06.03.2019 zielt darauf ab, eine Steuer von 3 % auf „digitale Umsätze“ zu erheben. Dies wurde rückwirkend zum 01.01.2019 eingeführt. Als digitale Umsätze zählen Umsätze, welche Einnahmen aus der Internetwerbung oder der Zurverfügungstellung einer digitalen Schnittstelle zum Datenaustausch generieren sowie der Verkauf von Nutzerdaten. Hierbei gelten als Besteuerungsbasis die Umsätze, welche Frankreich zugeteilt werden können. Dieser Anteil kann als Verhältnis der französischen Nutzer zu den weltweiten Nutzern ermittelt werden. Jedoch sind der französischen Digitalsteuer lediglich Konzerne unterworfen, die weltweit mindestens einen Umsatz von 750 Mio. Euro und zudem in Frankreich einen Umsatz von 25 Mio. Euro aufweisen. Laut Schätzungen sollen von dieser Digitalsteuer rund 20 bis 30 Unternehmen betroffen sein und Steuereinnahmen von ca. 500 Mio. Euro generieren. Frankreich ist jedoch sehr wohl bewusst, dass diese unilaterale Massnahme befristet ist.[156]

[146] Kempelmann, „The Future of International Tax,“ 664-665.

[147] Bendlinger, „BEPS 2.0 - Besteuerung digitalisierter Geschäfte - eine (un)endliche Geschichte,“ 108.

[148] Benjamin Grassl und Reinald Koch, „Unilaterale Initiativen zur Einführung von Digitalsteuern,“ IStR 28, Nr. 22 (2019): 873-874.

[149] Bendlinger, „BEPS 2.0 - Besteuerung digitalisierter Geschäfte - eine (un)endliche Geschichte,“ 108.

[150] Rat der EU für Wirtschaft und Finanzen

[151] Benedikt Mahr, „Digitalsteuer - kein Spielfeld für nationale Alleingänge,“ GWR, Nr. 11 (2019): 189.

[152] Grassl und Koch, „Unilaterale Initiativen zur Einführung von Digitalsteuern,“ 887.

[153] Zinowsky, „Abkehr von tradierten Verrechnungspreismethoden und Einführung einer weltweiten Mindestbesteuerung,“ 820.

[154] Boehle und Scholtholt, „Gestaltungsoptionen zur gerechten Besteuerung von Digitalkonzernen,“ 924.

[155] Grassl und Koch, „Unilaterale Initiativen zur Einführung von Digitalsteuern,“ 873.

[156] Jean-Paul Stévenard, „Der Entwurf einer französischen Digitalsteuer,“ IWB, Nr. 7 (2019): 270-271.

Nationaler Alleingang: Österreich
Österreich legte am 04.04.2019 einen Entwurf für eine österreichische Digitalsteuer vor, welcher seit 01.01.2020 in Kraft ist. Österreich bezieht sich mit dieser Digitalsteuer lediglich auf Online-Werbeleistungen, da derzeit eine Ungleichberechtigung zwischen digitaler und traditioneller Werbung herrscht. Seit 2000 wird in Österreich eine 5 %-ige Werbeabgabe auf Werbung in Print, Rundfunk, Fernsehen, Plakaten usw. erhoben. Digitale Werbung wird folglich davon nicht erfasst. Österreich soll mit dieser Steuer rund 30 Mio. Euro Steuereinnahmen im Jahr generieren. Wie auch in Frankreich sind lediglich Unternehmen von dieser österreichischen Digitalsteuer betroffen, die weltweit Umsätze von 750 Mio. Euro und in Österreich 25 Mio. Euro Umsätze verzeichnen. Die Höhe der Steuer beläuft sich auf 5 %, um die Gleichberechtigung der Werbeabgabe für Print-Medien zu gewährleisten und ist somit um 2 % höher als die von der EU vorgeschlagene Richtlinie. Sie unterscheidet sich somit auch von der französischen Digitalsteuer.[157]

3.2. Gewinnverteilungsregeln des Unified Approach

Die OECD arbeitet auf Hochtouren, um das internationale Steuerrecht wieder zu stabilisieren und so schnell wie möglich eine konsensbasierte Lösung zu entwickeln. Der neueste Entwurf des Unified Approach soll die Steuerregeln an die digitalisierte Zeit anpassen und auch für digitale Geschäftsmodelle anwendbar und fair sein. Laut OECD werden nun mehr Besteuerungsrechte in die Marktjurisdiktionen verlegt. Dies kann bei digitalen Geschäftsmodellen auch dazu führen, dass die Besteuerungsrechte den Staaten, in denen sich die User befinden, zugeordnet werden. In diesem Ansatz ist vorgesehen, drei verschiedene Arten des steuerpflichtigen Gewinnes, welche den Marktstaaten zugewiesen werden können, zu ermitteln. Diese Arten werden nachfolgend als Betrag A, Betrag B und Betrag C deklariert und nun ausführlicher und detaillierter beschrieben. Der Fokus dieser Analyse liegt bei Betrag A, da dieser die Neuheit unter diesen drei Beträgen darstellt.[158]

Die OECD erwähnt, dass alle diese Beträge einen unterschiedlichen Anwendungsbereich aufweisen und somit auch Überschneidungen möglich sind. Weiters werden lediglich unter Betrag A neue Nexus-Regeln verwendet, wobei die Beträge B und C immer noch auf den jetzigen Gewinnverteilungsregeln basieren, um Besteuerungsrechte zu verteilen. Dies bedeutet folglich auch, dass die Beträge B und C immer noch auf dem Prinzip der physischen Präsenz basieren, um die Praktikabilität der Verrechnungspreise weiter zu garantieren. Insbesondere der Betrag A revolutioniert das internationale Steuerrecht, indem das Konzept der physischen Präsenz verworfen wird und dieser Betrag A durch einen Formel-Ansatz berechnet wird.[159] Das Anknüpfen an eine nicht-physische Betriebsstätte erfordert indes neue Gewinnverteilungsregeln. Dies resultiert laut Schuster aus der Tatsache, dass die bisherigen Zuteilungsregeln auf Substanz, Funktionsausübung, Risikoübernahme und dem Fremdvergleichsgrundsatz basierten und bei nicht-körperlichen Anknüpfungspunkten somit unzureichend sind.[160]

Da es von wesentlicher Bedeutung ist, die Begriffe Residualgewinn und Routinegewinn zu verstehen, werden diese nun nachfolgend kurz erklärt. Diese Unterscheidung ist laut Schön bereits aus den Verrechnungspreisregeln bekannt, wird nun aber auch für den neuen Ansatz der OECD herangezogen und bei der Zuordnung von Gewinnen an Marktstaaten berücksichtigt.[161]

Routinegewinne werden, wie oben erwähnt, bereits in der Verrechnungspreisrichtlinie[162] der OECD für standardisierte Routinefunktionen verwendet, können jedoch auch auf alle Unternehmensbereiche angewendet werden. Dies bedeutet, dass allen Produktionsstandorten eines Unternehmens ein standardisierter fixer Gewinnanteil zugerechnet wird. Dieser wird abhängig von bestimmten Faktoren berechnet. Darunter fallen beispielsweise die Vermögenslage, der Personaleinsatz, der Geschäftszweck oder das Geschäftsrisiko dieses Produktionsstandortes. Oftmals gibt es laut Schön eine Differenz zwischen der Summe der Routinegewinne und dem Unternehmensgesamtgewinn. Diese Differenz wird als Residualgewinn bezeichnet. Dies kann auf unterschiedliche Faktoren, welche bei der Bestimmung des Routinegewinnes nicht miteinbezogen werden, zurückgeführt werden. Als Beispiel können Synergieeffekte im Konzern oder auch die Prämie auf das geschäftliche Gesamtrisiko des Konzernes genannt werden. Laut Schön wurde von der Wissenschaft bereits seit einigen Jahren gefordert, den Residualgewinn, ganz oder auch nur teilweise, den Marktstaaten zuzuteilen.[163] Zum jetzigen Zeitpunkt werden also zuerst die Funktionen, die laut Risiko- und Funktionsanalyse über wenig Risiko verfügen, mit dem Routinegewinn vergütet. Danach wird den Strategieträgern eines Konzernes, die das tatsächliche volle Risiko tragen, der Residualgewinn zugeteilt.[164]

3.2.1. Betrag A

Wie bereits erwähnt wurde, schafft der Betrag A ein neues Besteuerungsrecht, welches Teile des Residualgewinnes den Marktstaaten zuteilen kann. Jedoch fällt nicht jedes Unternehmen in diesen Anwendungsbereich, sondern lediglich

[157] Grassl und Koch, „Unilaterale Initiativen zur Einführung von Digitalsteuern," 884-886.
[158] OECD, „Outline of the Architecture of a Unified Approach on Pillar One," 8.
[159] Ebd., 8-9.
[160] Schuster, „Der "Unified Approach" der OECD - ein Überblick und erste Einschätzungen," 515.
[161] Schön, „Seminar F (IFA/OECD): Internationale Besteuerung der digitalisierten Wirtschaft," 649.
[162] OECD, OECD-Verrechnungspreisleitlinien für multinationale Unternehmen und Steuerverwaltungen 2017 (Paris: OECD Publishing, 2018).
[163] Schön, „Seminar F (IFA/OECD): Internationale Besteuerung der digitalisierten Wirtschaft," 649.
[164] Veronika Daurer, „Der Unified Approach der OECD," ÖStZ, Nr. 20 (2019): 555.

MNE, welche sowohl den neuen steuerlichen Anknüpfungspunkt als auch eine gewisse Höhe des Betrages A aufweisen. Für die OECD ist es von grosser Relevanz, dass dieser neue Ansatz sowohl eine Vereinfachung und Stabilisierung des Steuersystems garantiert, die Komptabilität mit den Verrechnungspreisgrundsätzen sicherstellt sowie auch in der Lage ist, Doppelbesteuerung effektiv zu vermeiden.[165] Diese neue Gewinnaufteilung folgt laut Daurer der Annahme, dass zumindest ein Teil der immateriellen Vermögenswerte, die den Residualgewinn eines Konzernes entstehen lassen, auch auf die Marktstaaten zurückzuführen sind.[166]

Anwendungsbereich und Carve-Outs

Für die OECD ist es eindeutig, dass Unternehmen auch ohne physische Präsenz einen wesentlichen Einfluss am Markt eines anderen Landes ausüben können, ohne in Infrastruktur oder Ähnliches investieren zu müssen. Der Betrag A soll an diesem Problem anknüpfen und somit einen Teil des Residualgewinnes den Marktstaaten zuteilen. In den Anwendungsbereich dieses Betrages fallen MNE, die mit oder auch ohne physische Präsenz wesentlich und nachhaltig am Wirtschaftsleben eines Staates Anteil haben.[167] Für die OECD ist es zudem wichtig, bestimmte Schranken einzuführen, sodass der Aufwand für Compliance und Administration in Relation der Vorteile angemessen bleibt.[168] Es werden nun die sechs verschiedenen Schwellen näher definiert.[169]

Schwelle 1: Von Betrag A werden nur die MNE betroffen sein, die eine gewisse Umsatzgrenze überschreiten. Die OECD sieht zum momentanen Zeitpunkt die Grenze von 750 Mio. Euro vor, welche ident wäre mit der Grenze, die für den Country-for-Country Report[170] nach BEPS-Aktionspunkt 13 ausschlaggebend ist. Damit sollen vor allem kleine Unternehmen geschützt werden, da für diese die Kosten zu hoch wären.

Schwelle 2: Die zweite Schwelle stellt einen Aktivitäten-Test dar. Hierbei sollen lediglich die Unternehmen in den Anwendungsbereich des Betrages A fallen, welche Aktivitäten aus Automated Digital Services bzw. Consumer-facing activities ausführen.

Schwelle 3: Falls ein Unternehmen die oben genannten Schranken übertritt, jedoch der total aggregierte Umsatz im Anwendungsbereich (also Umsätze aus Automated digital services und/oder Consumer-facing businesses) einen bestimmten de minimis-Schwellenwert nicht übersteigt, fällt dieses Unternehmen nicht in den Anwendungsbereich des Betrages A. Als Beispiel für einen Carve-Out aufgrund dieser Schwelle wird ein Unternehmen genannt, welches gross ist, sich jedoch vollkommen dem inländischen Markt widmet und fast kein Einkommen aus dem Ausland verzeichnet. Die Höhe dieser Schwelle ist von der OECD noch nicht festgesetzt.

Schwelle 4: Weiters werden für den Betrag A nur die Unternehmen berücksichtigt, welche mit ihren Geschäftsbereichen eine Rentabilität erreichen, die höher als ein bestimmter Prozentsatz ist.

Schwelle 5: Zudem wird eine weitere de minimis-Grenze errichtet, damit nur Unternehmen, bei welchen die Summe der Residualgewinne eine gewisse Grenze überschreitet, in den Anwendungsbereich fallen.

Schwelle 6: Zum Schluss muss noch geprüft werden, ob der neue Nexus in einem Staat erfüllt wird. Erst dann kann eine Marktjurisdiktion den Gewinn unter Betrag A eines Unternehmens besteuern. Hat das Unternehmen alle oben genannten Schwellen übertreten, fällt es in den Anwendungsbereich des Betrages A, welcher anschliessend berechnet und in die verschiedenen Marktjurisdiktionen verteilt wird.

Die untenstehende Abbildung 11 visualisiert die oben genannten Schwellen, die ein Unternehmen überschreiten muss, um in den Anwendungsbereich des Betrages A zu fallen.

In den Anwendungsbereich fallen für die OECD zwei wesentliche Gruppen von Unternehmen, die Automated digital services und die bereits im ersten Entwurf des Unified Approach erwähnten Consumer-facing businesses. Diese zwei Gruppen im Anwendungsbereich werden nun definiert und näher betrachtet.[172] Für die OECD fallen insbesondere diese Unternehmen in den Anwendungsbereich, da die Kundeninteraktion ohne Aufweisen einer physischen Präsenz immer einfacher wird. Weiters wird vor allem durch diese intensive Kundeninteraktion der Marktwert und der Umsatz dieser im Anwendungsbereich liegenden Unternehmen gesteigert.[173] Ein wesentlicher Unterschied des neuen und aktuellen Entwurfs des Unified Approach zum ersten Entwurf vom Oktober 2019 zeigt sich bei den Gruppen im Anwendungsbereich. Zum damaligen Zeitpunkt fielen Unternehmen, die durch die Bereitstellung digitaler Inhalte, welche zudem eine Consumer-facing Komponente aufwiesen, in den Anwendungsbereich. Der aktuelle Entwurf der OECD entkoppelt jedoch nun die Anforderung einer Consumer-facing Komponente in Bezug auf digitale Services. Dies bedeutet, dass die

[165] OECD, „Outline of the Architecture of a Unified Approach on Pillar One," 9.
[166] Daurer, „Der Unified Approach der OECD," 555.
[167] OECD, „Outline of the Architecture of a Unified Approach on Pillar One," 9.
[168] Ebd., 12.
[169] Ebd., 9-10, 26.
[170] Verpflichtung zur länderbezogenen Berichterstattung
[171] OECD, „Outline of the Architecture of a Unified Approach on Pillar One," 26.
[172] Ebd., 9.
[173] Ebd., 10.

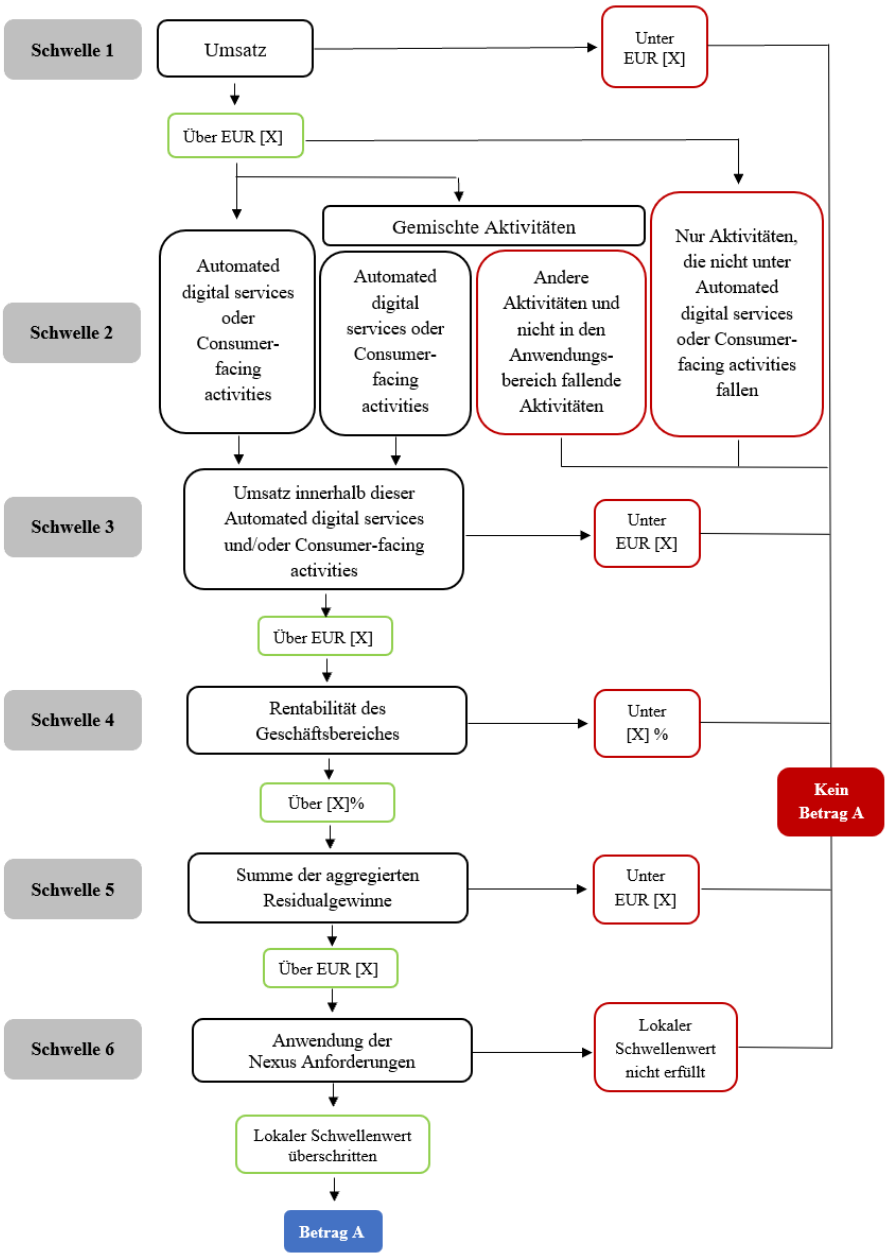


Abbildung 11: Begründung eines Nexus unter Betrag A
Eigene Abbildung in Anlehnung an OECD[171]

Gruppe der Automated digital services nun nicht mehr auf einer Consumer-facing Komponente basieren muss, um von Betrag A betroffen zu sein.[174]

Automated digital services

Für die OECD zeichnen sich diese Automated digital services vor allem dadurch aus, dass sie in der Lage sind, Dienste aus der Entfernung einer globalen und grossen Nutzerbasis anzubieten und dabei nicht über Infrastruktur in diesem Land zu verfügen. Zudem profitieren diese Unternehmen meist von Netzwerkeffekten und können mithilfe der User und deren Interaktionen einen wesentlichen Wert für das Unternehmen generieren. Damit sind vor allem Geschäftsmodelle gemeint, welche wesentlich von den Beiträgen der Nutzer profitieren und auch relevante Daten durch die Überwachung der Nutzer generieren können. In einigen dieser Unternehmen besteht ständige Interaktion zwischen dem Unternehmen und dem User, welcher ununterbrochen die Dienstleistungen des Unternehmens nutzen kann. Die Aufzählung der

[174]Vgl. OECD, „Secretariat Proposal for a “Unified Approach” under Pillar One,“ 7.

OECD beinhaltet somit Unternehmen, welche als Dienstleistung automatisierte digitale Medien bereitstellen und auch eine grosse Zahl an Kunden und Benutzern innerhalb mehrerer Jurisdiktionen aufweisen. Die OECD weist aber darauf hin, dass die untenstehende Aufzählung nicht abschliessend ist.[175]

Die OECD nennt verschiedene Beispiele von Unternehmen, die in diese Kategorie fallen:

- Social-Media Plattformen
- Online-Suchmaschinen
- Streaming digitaler Inhalte
- Online- Spiele
- Cloud-Computing Dienste
- Online-Vermittlungsplattformen, einschliesslich Online-Marktplätze
- Online-Werbedienste

Consumer-facing businesses

Die zweite Gruppe von Unternehmen, die von dieser neuen Regelung betroffen sind, sind die so genannten Consumer-facing businesses. Darunter fallen beispielsweise Unternehmen, die Einnahmen durch den Kauf und Verkauf von Waren und Dienstleistungen an Verbraucher generieren. Dabei ist unwesentlich, ob dies direkt oder indirekt stattfindet. Auch die Unternehmen, welche indirekt Produkte durch Intermediäre und Zwischenhändler, welche nur Routinetätigkeiten[176] ausführen, verkaufen, sind von dieser Regelung betroffen. Ausgenommen sind hierbei aber Unternehmen, die Zwischenprodukte und Komponenten vertreiben, welche in ein fertiges Produkt eingearbeitet und dann an Konsumenten verkauft werden. Weiters sollen auch Unternehmen, welche Lizenzeinnahmen aufweisen oder beispielsweise durch ein Franchisemodell Einnahmen erzielen, dieser Gruppe zugeordnet werden.[177]

Diese zweite Kategorie betrifft viele unterschiedliche Unternehmen, wobei auch traditionelle Unternehmen, die einen gewissen Grad an Digitalisierung aufweisen, darunterfallen. Immer mehr Unternehmen verwenden digitale Medien, um eine intensivere Kundeninteraktion und Kundenverbindung zu erreichen. Digitale Medien helfen dabei, Kundenbeziehungen durch gezieltes Marketing aufzubauen sowie Nutzerdaten zu erheben und gezielt zu nutzen. Dies scheint vor allem auf Unternehmen zuzutreffen, welche vernetzte Produkte verkaufen sowie hauptsächlich Online-Plattformen für den Verkauf und das Marketing nutzen.[178]

Die OECD listet folgende Beispiele für Consumer-facing businesses auf:

- Kleidung, Kosmetika, Toilettenartikel und Luxusgüter
- Personal Computer Produkte wie Mobiltelefone, Software und Haushaltsgeräte
- Markennahrungsmittel und Erfrischungsgetränke
- Franchisemodelle wie Lizenzvereinbarungen für den Restaurant- und Hotelsektor
- Automobile

Ausnahmen/Carve Outs

Die OECD nennt zudem verschiedene Ausnahmen, welche nicht von der neuen Regelung betroffen sein werden. Darunter fallen unter anderem die Rohstoffindustrie sowie auch alle anderen Hersteller und Verkäufer von Rohstoffen. Obwohl diese zu Produkten, welche schlussendlich an Konsumenten verkauft werden, verarbeitet werden, sind sie nicht in den Anwendungsbereich miteinzubeziehen. Während der Verkauf von rohen Kaffeebohnen ausserhalb des Anwendungsbereiches liegt, ist der Verkauf eines gebrandeten Glases Kaffee innerhalb des Anwendungsbereiches.[179] Greil und Wargowske begründen diese Ausnahme unter anderem damit, dass sehr viele Entwicklungsländer, die Rohstoffe exportieren, einen wichtigen Teil dieses Projektes der OECD darstellen und ansonsten negativ von diesem Ansatz betroffen wären.[180]

Als weitere Ausnahme wird von der OECD der Sektor rund um Finanzdienstleistungen genannt, da dieser Sektor weitestgehend mit gewerblichen Kunden interagiert. Darunter zählen auch Versicherungsdienstleister. Weiters werden auch die Luft- und Seefahrtunternehmen von dieser neuen Regelung ausgenommen. Dies wird dadurch begründet, dass in fast allen bilateralen Abkommen (DBA) das Besteuerungsrecht der Einkünfte dieser Unternehmen beim Wohnsitzstaat liegt. Diese gängige Praxis soll laut OECD nicht verändert werden, weshalb diese Unternehmen ausserhalb des Anwendungsbereiches sind.[181]

Es gibt viele Unternehmen, die mehrere Bereiche innerhalb des Unternehmens haben. Diese Unternehmen fallen mit manchen Aktivitäten in den Anwendungsbereich, während sie mit anderen ausserhalb des Anwendungsbereiches liegen. Hierbei gilt es abzuklären, inwiefern eine Segmentierung dieser unterschiedlichen Aktivitäten in verschiedene Geschäftsfelder möglich ist.[182] Für Daurer ist nicht absehbar, wie eine Segmentierung von Konzernergebnissen erstellt und geprüft werden kann, ohne einen zu hohen Aufwand dafür zu investieren.[183] Müller und Bauerfeld machen zudem darauf

[175] OECD, „Outline of the Architecture of a Unified Approach on Pillar One," 10.
[176] Beispiele der OECD: kleine Montage- und Verpackungsarbeiten
[177] OECD, „Outline of the Architecture of a Unified Approach on Pillar One," 10-11.
[178] Ebd., 9-11.
[179] Ebd., 11-12.
[180] Stefan Greil und Lars Wargowske, „Pillar 1 of the Inclusive Framework's Work Programme: The Effect on the Taxation of the Digital Economy and Reallocation of Taxing Rights," BIT 73, Nr. 10 (2019): 506.
[181] OECD, „Outline of the Architecture of a Unified Approach on Pillar One," 11-12.
[182] Ebd., 33.
[183] Daurer, „Der Unified Approach der OECD," 556.

aufmerksam, dass die Abgrenzung in verschiedene Geschäftsbereiche zu Unklarheiten und Unterschieden bei der Definition führen kann. Hierbei können Staaten ihre Definition der digitalen Wirtschaft ausweiten, um sich somit Besteuerungssubstrat zu sichern.[184] Auch Greil und Wargowske sind der Meinung, dass eine nötige Segmentierung von Geschäftsfeldern zu BEPS-Problemen führen kann, da Unternehmen ihre Gewinne in bestimmten Geschäftsfeldern anfallen lassen und somit die Bemessungsgrundlage senken können.[185]

Vor allem Unternehmen, welche hohe Umsätze ausserhalb des Anwendungsbereiches haben, müssen in diese unterschiedlichen Geschäftsfelder aufgeteilt werden, sodass nur Umsätze innerhalb des Anwendungsbereiches für den Betrag A berücksichtigt werden. Insbesondere auch für Unternehmen, die eine unterschiedliche Rentabilität in verschiedenen Ländern und Geschäftsbereichen haben, ist eine Segmentierung von Bedeutung. Auch erwähnt die OECD, dass Compliance-Kosten und die Administrierbarkeit und Genauigkeit gewährleistet werden muss.[186] Für Becker und Van der Ham wäre es sehr wohl möglich, Unternehmen mit gemischten Geschäftsmodellen die Entscheidung zu überlassen, ob diese nach vorhandenen oder nach den neuen Regeln besteuert werden sollen.[187]

Des Weiteren weist die OECD darauf hin, dass es viele Güter und Dienstleistungen gibt, die an Konsumenten und auch an Unternehmen verkauft werden. Falls diese Güter und Dienstleistungen jedoch gewöhnlich an Konsumenten verkauft werden, fallen auch diese Unternehmen in den Anwendungsbereich. Als Beispiel dafür wird ein Unternehmen genannt, welches Personal Computer an kleinere Unternehmen und Konsumenten verkauft.[188]

Nexus-Regeln

Des Weiteren werden neue steuerliche Anknüpfungspunkte für Unternehmen geschaffen, welche im Anwendungsbereich des Betrages A liegen. Die Nexus-Regeln werden darauf ausgelegt, inwiefern ein Unternehmen eine signifikante und nachhaltige Rolle in den Marktjurisdiktionen spielt. Die neuen Regeln sollen auch dazu führen, dass alle anderen steuerlichen Verpflichtungen in Verbindung mit der Allokation der Steuerrechte auf ein Minimum gekürzt werden. Für die OECD ist ein Indiz für eine signifikante und nachhaltige Rolle eines Unternehmens in einem Marktstaat die Generierung von Umsätzen im Anwendungsbereich über einige Jahre. Auch hier soll ein gewisser Schwellenwert, abhängig von der Grösse des Marktstaates, eingeführt werden. Diese Umsatzschwellen sollen im weiteren Verlauf noch festgesetzt werden.[189] Schuster macht darauf aufmerksam, dass seitens der OECD noch nicht geklärt wurde, inwieweit eine Marktjurisdiktion definiert werden kann. Es scheint durchaus möglich, dass aufgrund der Wesentlichkeit einer Region auch regionale Schwellenwerte einzuführen sind. Dies führe nicht nur zu unterschiedlichen Schwellenwerten der Staaten, sondern auch zu unterschiedlichen regionalen Schwellenwerten und würde ein „mögliches steuerliches Patchwork“ bedeuten. Dies kann auch Auswirkungen auf das Angebot derartiger Produkte haben, so Schuster.[190]

Für Unternehmen, die Automated digital services bereitstellen, reicht die oben genannte Umsatzschwelle, um in den Anwendungsbereich zu fallen und somit einen Betrag A nachzuweisen. Die OECD begründet dies durch die Tatsache, dass die Eigenschaften der Digitalisierung und der digitalen Geschäftsmodelle dazu führen, dass die Bereitstellung dieser Automated digital services eigentlich immer mit einer nachhaltigen Kundenbeziehung in Verbindung gebracht werden kann.[191] Daurer erwähnt, dass sich Aktivitäten wie Online-Werbung, die sich an die nicht zahlenden User richten, die in einem anderen Staat als die zahlenden User lokalisiert sind, auch in der Umsatzschwelle wiederfinden.[192]

Für die anderen Aktivitäten im Anwendungsbereich, also die Consumer-facing businesses, soll laut OECD nicht nur der Verkauf eines Produktes den neuen Nexus begründen. Vielmehr soll darauf abgezielt werden, dass Unternehmen in den Anwendungsbereich fallen, die explizit eine nachhaltige Kundeninteraktion in der Marktjurisdiktion aufweisen. Es soll daher nicht Unternehmen treffen, die Konsumgüter in einem Marktstaat verkaufen, ohne eine relevante oder nachhaltige Interaktion in diesem Staat zu verzeichnen. Hierbei sind laut OECD weitere Arbeiten nötig, um diese „Plus“-Faktoren zu identifizieren. Als Beispiel dafür könnte die gezielte Werbung in Bezug zu dieser Marktjurisdiktion genannt werden.[193]

Die OECD macht vor allem mehrmals darauf aufmerksam, dass diese neuen Nexus-Regeln vollkommen unabhängig existieren und keinen Effekt auf andere Steuer- und nicht Steuerregeln haben. Sie sollen keine „Spill-Over Effekte“ auf andere Nexus-Regeln haben. Dieser neue Nexus wird lediglich im Zusammenhang mit diesem neuen vorgeschlagenen Steuerrecht angewendet.[194] Weiters muss darauf hingewiesen werden, dass dieser neue Nexus auch durch den traditionellen Nexus begründet werden kann, also wie

[184] Sergej Müller und Tim Bauerfeld, „Risiken der Besteuerung der digitalen Wirtschaft nach „Säule 1“: Handlungsbedarf des deutschen Gesetzgebers,“ IWB, Nr. 23 (2019): 947.

[185] Greil und Wargowske, „Pillar 1 of the Inclusive Framework's Work Programme: The Effect on the Taxation of the Digi-tal Economy and Reallocation of Taxing Rights,“ 509.

[186] OECD, „Outline of the Architecture of a Unified Approach on Pillar One,“ 14.

[187] Jan D. Becker und Susann Van der Ham, „Das neue (digitale) ABC der Verlagerung von Besteuerungsrechten in die Marktstaaten: Neues OECD Konsultationspapier für die Weiterentwicklung der internationalen Unternehmensbesteuerung,“ DB 72, Nr. 46 (2019): 2542.

[188] OECD, „Outline of the Architecture of a Unified Approach on Pillar One,“ 12.

[189] Ebd., 12-13.

[190] Schuster, „Der “Unified Approach” der OECD - ein Überblick und erste Einschätzungen,“ 519.

[191] OECD, „Outline of the Architecture of a Unified Approach on Pillar One,“ 13.

[192] Veronika Daurer, „Unified Approach und Rechnungslegungsvorschriften - innovative Ideen für die Bemessung einer Ertragsteuer,“ RWZ 10, Nr. 2 (2020): 52.

[193] OECD, „Outline of the Architecture of a Unified Approach on Pillar One,“ 13.

[194] Ebd., 12-13.

bisher beispielsweise durch eine Betriebsstätte oder Tochtergesellschaft im Marktstaat.[195] Schuster erwähnt, dass die neuen Nexus-Regeln neben dem traditionellen Betriebsstättenbegriff existieren und ihn nicht ablösen sollen und somit eine neue Bestimmung in den DBA-Richtlinien notwendig ist. Jedoch kann das Existieren beider Bestimmungen laut Schuster wiederum zu Doppelbesteuerung führen.[196]

Höhe des Betrages A

Da nun erläutert wurde, wer von diesem neuen und revolutionären Betrag A betroffen ist und in den Anwendungsbereich fällt, ist der nächste Schritt die Höhe des Betrages A zu ermitteln. Der Betrag A wird mithilfe einer Formel berechnet, um den Betrag zu ermitteln, welcher als Teil der Residualgewinne den Marktjurisdiktionen zugeteilt werden kann. Wie aus unterstehender Abbildung 12 ersichtlich ist, handelt es sich bei Betrag A nur um den Teilbetrag, welcher ein bestimmtes Rentabilitätsniveau übersteigt.[197]

Zinowsky deutet darauf hin, dass sich die Rentabilität von Geschäftsaktivitäten wesentlich unterscheiden kann, wie beispielsweise durch regionale oder produktspezifische Unterschiede. Aus diesem Grund muss der Fokus der noch ausstehenden Arbeit auch darin liegen, neben der Praktikabilität auch die Varianz der Rentabilität zu berücksichtigen.[199]

Um zu garantieren, dass Routinegewinne von diesem Betrag A ausgeschlossen werden, wird es einen branchenabhängigen Prozentsatz geben, welcher abzuziehen sein wird. In einer Formel veranschaulicht bedeutet dies:[200]

$$z\% - x\% = y\% \tag{1}$$

z % Gewinnspanne (Gewinn : Umsatz) in Prozent
x % Routineanteil in Prozent
y % zu verteilender Residualgewinn aus Nicht-Routinetätigkeiten in Prozent

Im nächsten Schritt muss jedoch noch eine Abgrenzung zwischen dem Residualgewinn, der auf die Wertschöpfung[201] entfällt, und dem Residualgewinn, welcher auf Faktoren wie Rechte an geistigem Eigentum, Risiko und Kapital entfällt, gemacht werden. Dies dient dazu, dass der Residualgewinn nicht nur durch die Umsätze dieser Marktjurisdiktionen dargestellt werden kann, sondern auch durch die oben genannten Faktoren beeinflusst wird. Dies wird durch v % abgebildet und stellt den Anteil des Residualgewinnes dar, welcher nicht auf den Umsatz zurückgeführt werden kann.[202] Als Beispiel dafür nennt die OECD im ersten Entwurf des Unified Approach ein Social-Media Unternehmen, welches sowohl durch die Datensammlung der User sowie einer wertvollen Marke Residualgewinne generiert, jedoch auch durch Software oder Algorithmen die Residualgewinne steigern kann.[203]

$$w\% + v\% = y\% \tag{2}$$

w % Anteil der Marktjurisdiktion am Residualgewinn in Prozent
v % nicht-umsatzorientierter Anteil am Residualgewinn in Prozent
y % zu verteilender Residualgewinn aus Nicht-Routinetätigkeiten in Prozent

w % ist nun der Betrag, welcher schlussendlich auf die unterschiedlichen Marktjurisdiktionen, mithilfe eines Verteilungsschlüssels, verteilt werden kann.[204] Der Routineanteil in Prozent (x %) sowie auch der Anteil der Marktjurisdiktionen am Residualgewinn in Prozent (w %) ist seitens der OECD noch nicht festgelegt. Von der OECD werden bei der Berechnung des Betrages A zur wirtschaftlichen Analyse und Folgenabschätzung unter Säule 1 für x % ein Prozentsatz von 10 % bzw. 20 % und für w % ein Prozentsatz von 20 % angenommen.[205] Diese Annahmen der OECD werden somit auch in der Fallbearbeitung unter Kapitel 4 als Leitprozentsätze verwendet.

Daurer erwähnt, dass der ermittelte Residualgewinn für Betrag A rein fiktiv ist und somit vom Residualgewinn des Unternehmens anhand des Fremdvergleichsgrundsatzes abweichen kann. Zudem berücksichtigt der Residualgewinn unter Betrag A die ganze Unternehmensgruppe und bezieht sich somit nicht auf einzelne Steuersubjekte.[206]

Der Betrag A soll auf einem konsolidierten Konzernabschluss basieren und unterscheidet sich somit von den bestehenden Verrechnungspreisregeln. Die OECD legt keinen bestimmten Rechnungslegungsstandard fest, da die Unterschiede zwischen diesen lediglich Zeitdifferenzen darstellen. Diese Zeitdifferenzen beeinflussen das aggregierte Einkommen jedoch nicht, weshalb dies irrelevant ist. Die Aufwände, um die verschiedenen Rechnungslegungsstandards der Jurisdiktionen zu harmonisieren, soll auf ein Minimum reduziert werden. Es wird lediglich an den Unterschieden gearbeitet, welche in der Höhe und Dauer erheblich sind.[207] Daurer

[195] Daurer, „Unified Approach und Rechnungslegungsvorschriften - innovative Ideen für die Bemessung einer Ertragsteuer," 53.
[196] Schuster, „Der "Unified Approach" der OECD - ein Überblick und erste Einschätzungen," 515.
[197] OECD, „Outline of the Architecture of a Unified Approach on Pillar One," 13-14.
[198] OECD, „Tax Challenges arising from the Digitalisation of the Economy," 11.
[199] Zinowsky, „Abkehr von tradierten Verrechnungspreismethoden und Einführung einer weltweiten Mindestbesteuerung," 814.
[200] Schuster, „Der "Unified Approach" der OECD - ein Überblick und erste Einschätzungen," 516.
[201] Konsumorientierte Wertschöpfung in den Marktstaaten
[202] Schuster, „Der "Unified Approach" der OECD - ein Überblick und erste Einschätzungen," 516; Becker und Van der Ham, „Das neue (digitale) ABC der Verlagerung von Besteuerungsrechten in die Marktstaaten," 2543.
[203] OECD, „Secretariat Proposal for a Ünified Approachünder Pillar One," 15.
[204] Schuster, „Der "Unified Approach" der OECD - ein Überblick und erste Einschätzungen," 516.
[205] OECD, „Tax Challenges arising from the Digitalisation of the Economy," 12.
[206] Daurer, „Unified Approach und Rechnungslegungsvorschriften - innovative Ideen für die Bemessung einer Ertragsteuer," 53.
[207] OECD, „Outline of the Architecture of a Unified Approach on Pillar One," 13-14.

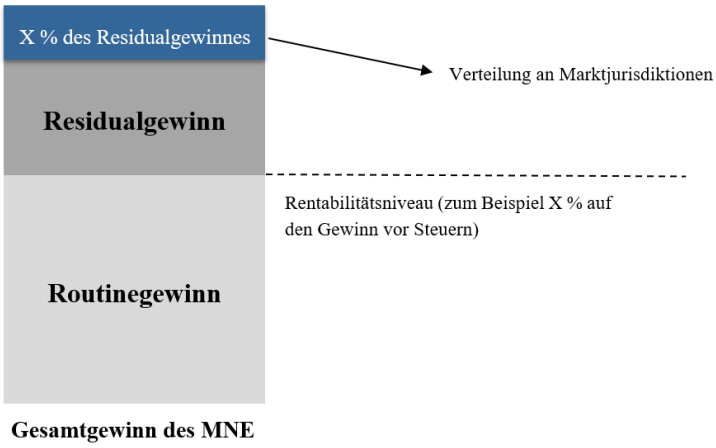


Abbildung 12: Höhe des Betrages A
Eigene Abbildung in Anlehnung an OECD[198]

geht davon aus, dass sich die OECD auf die wichtigsten permanenten Differenzen, die durch Unterschiede der steuerlichen und unternehmensrechtlichen Gewinnermittlung entstehen, konzentrieren wird. Als Beispiel dafür können Abzugsverbote angesehen werden. Die wirkliche Schwierigkeit liegt laut Daurer aber darin, dass dieser Ansatz global verwirklicht werden muss und dementsprechend auch die „universellen Besonderheiten der steuerlichen Gewinnermittlung identifiziert“ werden müssen. Temporäre Differenzen, welche sich über die Jahre wieder ausgleichen und nur zu einer zeitlichen Verschiebung des Steuersubstrats nach vorne bzw. hinten führen, wird man laut Daurer nicht berücksichtigen. Als Beispiel dafür können Schwankungen der Bilanzwerte und daraus entstehende Schwankungen auf den Gewinn genannt werden. Diese Schwankungen sind beispielsweise durch eine Bewertung nach International Financial Reporting Standards (IFRS) zum Fair Value- Marktwert meist grösser als durch eine Bewertung nach dem Unternehmensgesetzbuch (UGB), welches zu Anschaffungskosten bewertet. Jedoch sind einige temporäre Differenzen langanhaltend, was dazu führen könnte, dass auch diese angepasst werden müssen.[208] Für Pistone et al. führen die Unterschiede in den Rechnungslegungsstandards der Jurisdiktionen zu einem Anstieg der Doppelbesteuerung. Die Bewertung des Gesamtgewinnes soll die Rechnungslegungsstandards, die von der Jurisdiktion, in der die Mutter des MNE ansässig ist, vorgegeben wird, verfolgen.[209] Die konsolidierten Konzernabschlüsse aus internationalen Rechnungslegungsstandards als Berechnungsbasis für den Residualgewinn einzusetzen, sieht auch Schuster als kritisch an. Einige Staaten, so auch Österreich, haben nationale Rechnungslegungsvorschriften, welche andere Ziele als die internationalen Rechnungslegungsstandards verfolgen. Somit können immense Unterschiede zwischen nationalen und internationalen Rechnungslegungsvorschriften entstehen. Dies kann zu grösseren Konflikten führen, falls eine Berichtigung der Verrechnungspreise gemacht werden muss, wie unter Betrag B und C geplant ist.[210] Ahlers hebt indes hervor, dass seit rund 20 Jahren innerhalb der EU eine einheitliche Bemessungsgrundlage für Körperschaftsteuern diskutiert wird, weshalb eine Lösungsfindung auf der globalen Ebene des Inclusive Frameworks noch unwahrscheinlicher scheint. Das Problem der harmonisierten Gewinnermittlung kann als Hauptproblem der Säule 1 angesehen werden, so Ahlers.[211]

Als Berechnungsbasis für den Betrag A wird von der OECD das Ergebnis vor Steuern anvisiert, da dies in den meisten Fällen auch die Basis für die nationale Körperschaftsteuer ist. Dieser Ansatz wird sowohl für Gewinne als auch für Verluste gelten und wird mit einer Vortragsregelung ausgestattet sein, so die OECD.[212] Für Zinowsky ist das sachgerechte Umgehen mit Verlusten höchst relevant, da der Ansässigkeitsstaat eines Unternehmens kaum damit einverstanden ist, wenn beim Gewinnfall Steuersubstrat abgegeben, beim Verlustfall der Steuerausfall jedoch alleine getragen werden muss.[213] Zudem soll mit einer Verlustvortragsregel die Benachteiligung von neu entstandenen Unternehmen wie Start-Ups verhindert werden.[214]

Für die OECD scheint es durchaus möglich, die Höhe des Betrages A vom Grad der Digitalisierung der Geschäftsaktivitäten abhängig zu machen. Dabei spricht die OECD von Digital Differentiation. Es sind bezüglich des Betrages A noch weitere Arbeiten erforderlich, um zu klären, wie hoch das

[208] Daurer, „Unified Approach und Rechnungslegungsvorschriften - innovative Ideen für die Bemessung einer Ertragsteuer,“ 54.
[209] Pistone et al., „The OECD Public Consultation Document “Secretariat Proposal for a ‘Unified Approach’ under Pillar One”: An Assessment,“ 23.
[210] Schuster, „Der “Unified Approach” der OECD - ein Überblick und erste Einschätzungen,“ 520.
[211] Henrik Ahlers, „Gründlichkeit vor Schnelligkeit,“ DB 72, Nr. 50 (2019): M4.
[212] OECD, „Outline of the Architecture of a Unified Approach on Pillar One,“ 13-14.
[213] Zinowsky, „Abkehr von tradierten Verrechnungspreismethoden und Einführung einer weltweiten Mindestbesteuerung,“ 813-814.
[214] Pistone et al., „The OECD Public Consultation Document “Secretariat Proposal for a ‘Unified Approach’ under Pillar One”: An Assessment,“ 21.

Rentabilitätsniveau sein soll und welcher Anteil der Residualgewinne den Marktjurisdiktionen zugeteilt werden soll.[215]

Verteilung des Betrages A

Nachdem die Höhe des Betrages A festgelegt worden ist, muss dieser Betrag A auf die verschiedenen Marktjurisdiktionen mithilfe eines Verteilungsschlüssels aufgeteilt werden. Dieser Zuordnungsschlüssel beruht auf den Umsätzen, welche den neuen Nexus unter Betrag A begründen.[216] Es ist noch offen, in welcher Form dieser Betrag A erhoben werden wird. Laut Boehle und Scholtholt ist eine Quellensteuer jedoch denkbar.[217] Für Daurer bleibt zudem noch zu klären, wie die Einhebung des Betrages A, insbesondere bei Fällen, wo keine physische Präsenz im Marktstaat vorhanden ist, gewährleistet werden kann.[218]

Laut Müller und Bauerfeld zeigt das untenstehende Beispiel in Abbildung 13, dass die Zuteilung des Residualergebnisses an die Marktstaaten nicht wirklich im Einklang mit der digitalen Wertschöpfung steht. Zudem zielt der Vorschlag von Säule 1 darauf ab, dass Routinegesellschaften nun mit einem Gewinn, der nicht übereinstimmend mit den eingesetzten Wirtschaftsgütern und den Risiken ist, vergütet werden.[219]

Unternehmen A, ansässig in Staat A, ist hierbei ein Suchmaschinenanbieter, der rund 100 Mitarbeiter beschäftigt, die an der Programmierung arbeiten. In Staat B hat dieses Unternehmen eine Vertriebseinheit mit lediglich zwei Mitarbeitern. Diese Einheit generiert jedoch sehr hohe Umsätze.

Laut des Ansatzes der Säule 1 muss nun ein Teil des Residualgewinnes dem Land B zugeordnet werden. Jedoch tragen die Mitarbeiter des Unternehmens B nur sehr gering bis gar nicht zur Wertschöpfung bei, da die Progammierarbeit, also die relevante Arbeit für das Geschäftsmodell, in Land A durchgeführt wird. Die Allokation eines Teiles des Residualgewinnes in Staat B scheint für Müller und Bauerfeld nicht im Einlang mit der digitalen Wertschöpfung zu stehen.[221]

Für Becker und Van der Ham scheint es vor allem fraglich, wie die Zuteilung der Umsätze bei indirekten Verkäufen über Grosshändler, die verschiedene Märkte abdecken, praktisch durchgeführt werden kann. Es stellt sich die Frage, ob die Umsätze dem Ansässigkeitsstaat des Zwischenhändlers oder den Staaten der Endkonsumenten, welche von diesem Zwischenhändler beliefert werden, zugerechnet werden. Zudem scheint fraglich, wie dies bei mehreren Zwischenstufen zwischen Endkonsument und Hersteller geregelt werden kann. Weiters heben die Autoren auch hervor, dass in manchen Branchen die Zuordnung anhand von Umsätzen zu immensen Schwierigkeiten führen kann. Hierbei nennen sie als Beispiel die Pharmaindustrie, in der sehr oft Re-Importe durchgeführt werden, wobei diese teilweise sogar regulatorisch festgelegt und gefordert werden. Diese Re-Importe sind in der Praxis nur schwer nachvollziehbar und können deshalb grosse Schwierigkeiten bei der Implementierung des Unified Approach nach sich ziehen.[222]

3.2.2. Betrag B

Der Betrag B soll eine feste Rendite für Basismarketing- und Vertriebsaktivitäten darstellen und fungiert somit als Fixbetrag für erbrachte grundlegende Marketing- und Vertriebsaktivitäten in den Marktstaaten. Dies bedeutet, dass Vertriebshändler, unabhängig ob als Tochterunternehmen oder Betriebsstätte gegründet, die Produkte kaufen und dann wieder verkaufen, davon betroffen sind. Diese pauschal ermittelte Routinevergütung für Vertriebsaktivitäten soll auf dem Fremdvergleichsgrundsatz basieren und ist laut OECD nicht optional. Dieser Betrag dient dazu, eine Vereinfachung für die Berechnung der pauschalen Gewinnsätze der Aktivitäten darzustellen und damit Ressourcen zu schonen. Er soll somit helfen, die Komplexität und Compliance-Kosten rund um Verrechnungspreisregeln zu senken sowie auch Steuersicherheit für die Preisgestaltung von Transaktionen gewährleisten, um Streitigkeiten zu minimieren. Die OECD hebt hervor, dass es wichtig ist, Basisvertriebs- und Marketingaktivitäten nur in Betrag B und nicht in Betrag C zu berücksichtigen. Dies muss mit einer klaren Definition garantiert werden.[223] Während der Betrag A an neue Nexus-Regeln anknüpft, basieren sowohl Betrag B als auch C immer noch auf dem Konzept der physischen Präsenz und sollen dabei helfen, Verrechnungspreisrichtlinien zu vereinfachen.[224]

Der Betrag B soll als Ersatz für Verrechnungspreisgrundsätze auf bestimmte Marketing- und Vertriebsaktivitäten gelten. Den Grund für die Ausarbeitung dieses Betrages als eine Fixabgabe nennen Boehle und Scholtholt darin, dass abkommensrechtlich die Besteuerungsrechte dieser Aktivitäten zwar bereits bei den Marktstaaten liegen, dies in der Vergangenheit jedoch zu unzähligen Konflikten führte.[225] Die fixe Vergütung könnte laut Daurer auch die Notwendigkeit der einzelfallbezogenen Sachverhalts- und Vergleichbarkeitsanalysen (wie Benchmarks) reduzieren und branchenabhängig entwickelt werden.[226]

Die OECD hat noch nicht genau definiert, was unter diesen Basisaktivitäten verstanden werden kann. Unter den Anwendungsbereich fallen wahrscheinlich Vertriebsvereinbarungen, welche ein standardmässiges Level von Funktionalität aufweisen. Dies bedeutet also kein Besitz von immateriellen Vermögenswerten sowie gar kein oder nur ein

[215] OECD, „Outline of the Architecture of a Unified Approach on Pillar One," 14.
[216] Ebd.
[217] Boehle und Scholtholt, „Gestaltungsoptionen zur gerechten Besteuerung von Digitalkonzernen," 922.
[218] Daurer, „Der Unified Approach der OECD," 556.
[219] Müller und Bauerfeld, „Risiken der Besteuerung der digitalen Wirtschaft nach „Säule 1"," 948.
[220] Ebd.
[221] Ebd.
[222] Becker und Van der Ham, „Das neue (digitale) ABC der Verlagerung von Besteuerungsrechten in die Marktstaaten," 2542.
[223] OECD, „Outline of the Architecture of a Unified Approach on Pillar One," 16.
[224] Daurer, „Der Unified Approach der OECD," 556.
[225] Boehle und Scholtholt, „Gestaltungsoptionen zur gerechten Besteuerung von Digitalkonzernen," 922.
[226] Daurer, „Der Unified Approach der OECD," 556.

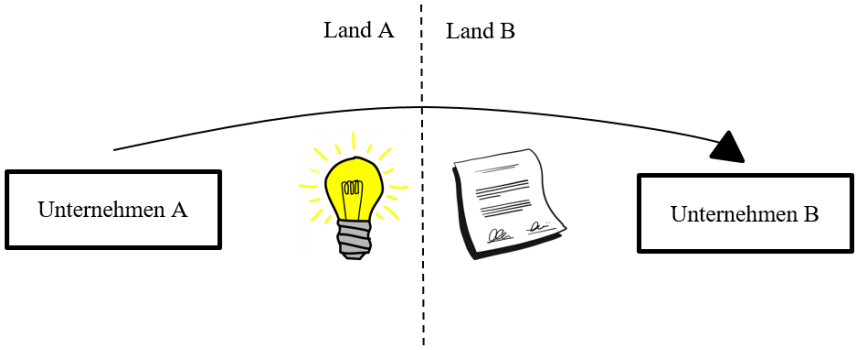


Abbildung 13: Zuteilung des Residualgewinnes an Routineunternehmen
Eigene Abbildung in Anlehnung an Müller und Bauerfeld[220]

geringes Risiko. Es soll ein Katalog entwickelt werden, in dem definiert werden kann, welche Aktivitäten und Unternehmen in den Anwendungsbereich fallen und welche nicht. Es sollen dazu auch die Verrechnungspreisregeln mancher Länder zur Hilfe gezogen werden. Für die Länder bedeutet die fixe Rendite, dass sie Kompromisse schliessen müssen, um sowohl die Einhaltung des Fremdvergleichsgrundsatzes sowie die einfache Administration von Betrag B zu gewährleisten. Die OECD hebt hervor, dass dieser Betrag B nicht alle Fälle individuell erfassen kann, jedoch der Vorteil in einer signifikanten Verbesserung und Vereinfachung zur Renditeermittlung der Aktivitäten im Anwendungsbereich liegt. Die Implementierung des Betrages B sollte nicht zu komplex sein, da keine Vertragsänderungen nötig sind. Da der Betrag B auf den Verrechnungspreisrichtlinien basiert, sind die bestehenden Verträge laut OECD ausreichend.[227]

Aktivitäten, die somit nicht in die Basisvertriebs- und Marketingaktivitäten eingeordnet werden können, sind beispielsweise Forschung und Entwicklung, Produktion und gruppeninterne Finanzdienstleistungen. Diese würden somit nicht einer festen Rendite unterliegen und werden weiterhin unter den bereits bestehenden Verrechnungspreisregeln abgewickelt.[228] Becker und Van der Ham machen darauf aufmerksam, dass dieser Vorschlag zur pauschalierten Vergütung der OECD an die Verrechnungspreisregeln in Brasilien erinnert. Hier wurden feste Margen für bestimmte Aktivitäten im Gesetz festgelegt. Diese bergen in der Praxis oft problematische Hürden, weshalb die Diskussionen um Betrag B noch nicht abgeschlossen sind. Diese pauschalierten Routinevergütungen vereinfachen dieses System sehr, weshalb dies auch Vorteile für die Steuerpflichtigen und die Finanzverwaltungen bringen würde. Es scheint jedoch fast unmöglich, einen Prozentsatz, welcher alle Unternehmen, Geschäftsmodelle und Staaten genügend berücksichtigt, festzusetzen.[229]

Während für Becker und Van der Ham eine pauschale Rendite zu mehr Komplexität führt, wird die Besteuerung für Pistone et al. dadurch vereinfacht. Die pauschale Rendite soll zu weniger Compliance- und Audit-Kosten führen und soll auch die Rechtssicherheit erhöhen. Eine fundierte wirtschaftliche Analyse der Märkte in regelmässigen Abschnitten ist nötig, um die jeweiligen Pauschalsätze zu definieren. Diese sind dann laut Pistone et al. aber sehr gut in der Lage, Konflikte und Streitpotential effektiv zu reduzieren.[230] PricewaterhouseCoopers (PwC) führte hierzu eine Analyse der Schwankungen der Gewinnmargen in verschiedenen Regionen und Industrien durch, da diese fixen Prozentsätze auch politisch von enormer Bedeutung sind. Hierbei wird ersichtlich, dass die Gewinnmargen der Unternehmen sehr unterschiedlich sein können. Beispielsweise liegt der Median der EBIT-Marge im Jahr 2017 in den USA bei 6,47 %, in Europa bei 4,39 % und in Asien bei 4,92 %. Weiters sind auch die Schwankungen innerhalb verschiedener Branchen sehr hoch. Während der Median der EBIT-Marge 2017 in der Immobilienbranche bei 23,6 % liegt, befindet sich der Median der EBIT-Marge in der Dienstleistungsbranche lediglich bei 3,1 %.[231]

3.2.3. Betrag C

Betrag C soll dazu dienen, Aktivitäten, die nicht oder zu wenig unter Betrag B berücksichtigt worden sind, zu erfassen und ergänzt somit Betrag B. Ein Konzern wird mit Betrag C beispielsweise steuerpflichtig, falls mehr Marketing- und Distributionsfunktionen durchgeführt wurden, als in Betrag B erfasst werden. Unter Betrag C ist es zudem notwendig, effektive Streitvermeidungs- und Beilegungsmechanismen zu etablieren.[232] Daurer erwähnt, dass dieser Betrag C aber nur dann eine Besteuerung des Marktstaates sichert, wenn sich dieser Marktstaat zu effektiven Streitbeilegungsmassnahmen verpflichtet, wie beispielsweise Schiedsverfahren.[233]

[227] OECD, „Outline of the Architecture of a Unified Approach on Pillar One," 17.
[228] PricewaterhouseCoopers, „Reaction to the OECD secretariat's consultation paper on the unified approach under Pillar 1," 24.
[229] Becker und Van der Ham, „Das neue (digitale) ABC der Verlagerung von Besteuerungsrechten in die Marktstaaten," 2544.
[230] Pistone et al., „The OECD Public Consultation Document "Secretariat Proposal for a 'Unified Approach' under Pillar One": An Assessment," 15.
[231] PricewaterhouseCoopers, „Reaction to the OECD secretariat's consultation paper on the unified approach under Pillar 1," 18-19.
[232] OECD, „Outline of the Architecture of a Unified Approach on Pillar One," 8.
[233] Daurer, „Der Unified Approach der OECD," 556.

Für Schuster sind die Beträge B und C nicht revolutionär, da sie immer noch auf dem Konzept der physischen Präsenz basieren. Jedoch ist die Berechnung neuartig. Der Betrag C scheint zudem besonders anfällig für Streitigkeiten, da dieser Betrag komplett einseitig von einem Staat gefordert werden kann.[234] Da laut OECD die Gewinnanteile unter Betrag A und die Festlegung des Betrages C interagieren, müssen effektive Mechanismen gegen Doppelbesteuerung geschaffen werden, sodass Gewinnanteile nicht doppelt besteuert werden.[235]

3.2.4. Safe-Harbour Ansatz für Säule 1

Des Weiteren wird von der OECD der Vorschlag des US-Treasury Sekretariats geprüft, Säule 1 als Safe-Harbour Ansatz einzuführen. Dies könnte dazu führen, dass sich ein MNE auf globaler Ebene freiwillig dazu bereit erklären kann, unter Säule 1 besteuert zu werden. Die OECD betont, dass noch viele Arbeiten nötig sind und dieser Vorschlag geprüft werden muss. Es sprechen jedoch viele Mitglieder des Inclusive Framework Bedenken dazu aus, da dieser die Steuersicherheit gefährde und damit die bisher definierten Anforderungen an Säule 1 nicht umgesetzt werden können. Die Entscheidung, ob Säule 1 als Safe-Harbour eingeführt wird oder nicht, ist laut OECD von grosser Bedeutung, um schlussendlich eine globale Lösung zu erreichen.[236] Das im Februar 2020 stattgefundene Update der OECD zur wirtschaftlichen Analyse und Folgenabschätzung zu Säule 1 deutet jedoch darauf hin, dass die Implementierung der Säule 1 als Safe-Harbour Ansatz eher unwahrscheinlich scheint.[237]

Schon in den eingelangten Kommentaren zum ersten Konsultationsdokument des Unified Approach wurde über einen möglichen Safe-Harbour Ansatz unter Betrag B diskutiert. Unter Betrag B bedeutet der Safe-Harbour Ansatz, dass von dieser fixen Rendite nach oben oder unten abgewichen werden kann. MNE müssten Gründe aufzeigen, damit sie einen grösseren oder auch kleineren Betrag B verlangen können. Falls mehr als die festgesetzte Rendite von einem MNE verlangt wird, könnte dies zu Überschneidungen mit dem Betrag C führen. Für PwC würde ein Safe-Harbour Ansatz unter Betrag B nur zu weiteren Konflikten und mehr Unsicherheit führen.[238] Deloitte hingegen argumentiert für den Safe-Harbour Ansatz. Für Unternehmen soll es möglich sein, eine niedrigere Rendite als die Basisrendite einzufordern, wenn dies auch anhand von Fremdvergleichsgrundsätzen angemessen scheint.[239]

3.2.5. Vermeidung von Doppelbesteuerung und Komplexität unter Säule 1

Die OECD macht zudem darauf aufmerksam, dass effektive Mechanismen zur Vermeidung von Doppelbesteuerung ermittelt werden müssen. Das Zusammenspiel der Beträge A und B ist laut OECD nicht kritisch in Bezug auf eine Doppelbesteuerung, da der Betrag A einen Teil der Residualgewinne an die Marktstaaten vergütet, während Betrag B eine fixe Renditevergütung für Basismarketing- und Vertriebsaktivitäten darstellt.[240]

Jedoch kann das Zusammenspiel von Betrag A und C zu Problemen bezüglich Doppelbesteuerung führen, so die OECD. Der Betrag A ist eine Überlagerung des Systems, dies bedeutet, dass der Gewinn eines Unternehmens bereits auf Grundlage der Verrechnungspreise zugewiesen wurde. Aufgrund der Tatsache, dass die Verrechnungspreise den ganzen Gewinn des MNE hineinbeziehen, müssen wirksame Aktivitäten gegen die Doppelbesteuerung umgesetzt werden. Eine Interaktion der Beträge A und C findet dann statt, wenn die Aktivitäten eines MNE sowohl von Betrag A erfasst werden, als auch einer Neubewertung der Verrechnungspreise unterworfen sind.[241] In der Praxis werden laut OECD zuerst die Gewinne eines MNE gemäss Verrechnungspreisen allokiert (somit Betrag B und C) und dann wird Betrag A gemäss den neuen Steuerregeln auf die Staaten verteilt.[242] Für Daurer ist die Tatsache, dass der Betrag A eine Überlagerung des Systems darstellt, als sehr kritisch anzusehen. Die Gewinne eines Unternehmens, die zuvor schon von den Verrechnungspreisregeln besteuert werden, nochmals zusätzlich zu besteuern, führt zu einer Doppelbesteuerung. Um dies zu vermeiden, soll laut Daurer der konkrete Steuerpflichtige des Betrages A ermittelt werden, sodass dieser den Betrag A bei seiner „traditionellen“ Gewinnermittlung gewinnmindernd berücksichtigen kann.[243] Der OECD ist sehr wohl bewusst, dass Streitigkeiten rund um Betrag A massive Folgen haben können, da damit oftmals zahlreiche Marktjurisdiktionen betroffen sind. Um dieses Szenario zu vermeiden, werden effektive Streitbeilegungsmassnahmen umgesetzt, so die OECD.[244]

Für Zinowsky wird durch die Säule 1 die Komplexität wesentlich erhöht. Es werden mehr Konflikte im Bereich der Verrechnungspreise entstehen, da in der Zukunft die Wertschöpfung nicht mehr vom Konzept der physischen Präsenz, angefallenen Kosten oder vorhandenen Funktionen abhängig gemacht wird. Die Compliance-Kosten der Unternehmen werden sich erhöhen. Für Unternehmen wird es nun von Bedeutung sein, dass sie in Zukunft nachweisen können, dass keine Wertschöpfungsbeiträge existieren.[245] Olbert und Sprengel weisen darauf hin, dass eine Anpassung der Verrechnungs-

[234] Schuster, „Der “Unified Approach” der OECD - ein Überblick und erste Einschätzungen,“ 517-520.
[235] OECD, „Outline of the Architecture of a Unified Approach on Pillar One,“ 16.
[236] Ebd., 4, 20-21.
[237] OECD, „Tax Challenges arising from the Digitalisation of the Economy,“ 8.
[238] PricewaterhouseCoopers, „Reaction to the OECD secretariat’s consultation paper on the unified approach under Pillar 1,“ 32.
[239] Deloitte, „Reaction to the OECD secretariat’s consultation paper on the unified approach under Pillar 1,“ 8.
[240] OECD, „Outline of the Architecture of a Unified Approach on Pillar One,“ 16.
[241] Ebd., 14-15.
[242] Ebd., 15.
[243] Daurer, „Der Unified Approach der OECD,“ 556.
[244] OECD, „Outline of the Architecture of a Unified Approach on Pillar One,“ 18.
[245] Zinowsky, „Abkehr von tradierten Verrechnungspreismethoden und Einführung einer weltweiten Mindestbesteuerung,“ 819.

preise an digitale Geschäftsmodelle sehr wohl machbar wäre und somit die Einführung eines neuen Nexus nicht notwendig ist. Die Autoren vertreten die Meinung, dass eine Anpassung der bestehenden Gewinnverteilungsregeln einen Grossteil der Probleme der digitalen Wirtschaft lösen kann.[246]

Auch Greil und Wargowske heben hervor, dass der Unified Approach im Zusammenhang mit den Fremdvergleichsgrundsätzen einige Schwierigkeiten bereiten wird, insbesondere wenn die Fremdvergleichsgrundsätze nicht in einem Zug mit dem neuen Besteuerungsrecht angepasst werden. Die Interaktion dieser beiden Ansätze scheint noch nicht eindeutig, so die Autoren.[247] Für Heidecke et al. ist die Abkehr vom Fremdvergleichsgrundsatz insoweit problematisch, da dieser im Moment den gemeinsamen Nenner darstellt, „anhand dessen die verschiedenen Jurisdiktionen die steuerliche Angemessenheit von konzerninternen Liefer- und Leistungsbeziehungen bemessen bzw. überprüfen können". Der Fremdvergleichsgrundsatz fungiert somit als Schnittstelle zwischen dem Steuerrecht und den innerbetrieblichen Geschäftsprozessen. Jedoch werden nun aufgrund politisch motivierter Willensäusserungen die Grundsätze des Fremdvergleichsgrundsatzes nach und nach untergraben.[248]

Boehle und Scholtholt erwähnen, dass, obwohl die Vermeidung der Doppelbesteuerung von grosser Relevanz für die OECD scheint, dies in den ausgearbeiteten Gewinnverteilungsregeln nicht sichtbar wird, da zur Vereinfachung in Streitfragen immer von einem Näherungsverfahren Gebrauch gemacht wird.[249] Becker und Van der Ham machen darauf aufmerksam, dass die Ausweitung von Besteuerungsrechten in die Marktstaaten, also eine Verdoppelung der Besteuerungsrechte, sicher auch die Fälle der Doppelbesteuerung wesentlich erhöhen wird.[250] Für Greil und Wargowske als auch für Fuest scheint es eindeutig, dass eine hohe Komplexität des Steuersystems auch zu einer grösseren Möglichkeit führt, Schlupflöcher in Hinsicht auf BEPS auszunutzen.[251] Auch Schuster ist der Meinung, dass dieser neue Ansatz eine weitere administrative Hürde für Unternehmen bergen wird. Obwohl die OECD dagegen ankämpft, muss darüber diskutiert werden, inwieweit die momentanen Pflichten für Unternehmen abgeschafft oder mindestens reduziert werden können. Die Compliance-Kosten der Unternehmen beanspruchen schon jetzt zu viele Unternehmensressourcen, so Schuster.[252] Für Daurer würde die Umsetzung des Unified Approach dem „bereits relativ komplexen System der internationalen Unternehmensbesteuerung eine weitere Ebene an Komplexität hinzufügen".[253]

3.2.6. Ausstehende Arbeiten für das Jahr 2020

Um im Jahr 2020 einen globalen Konsens über Säule 1 zu erreichen, bedarf es noch einer Klärung vieler technischer Fragestellungen. Hierbei bestehen innerhalb des Betrages A immer noch zahlreiche Unklarheiten, die geklärt werden müssen, um Steuersicherheit zu garantieren. Dies ist vor allem der Anwendungsbereich rund um Betrag A. Dabei gilt es, die Definitionen der betroffenen Unternehmen weiter zu detaillieren sowie die Schwellenwerte mit passenden Zahlen und Prozentwerten zu definieren. Weiters müssen auch die Ausnahmen der Säule 1 weiter und detaillierter eruiert werden. Zudem muss das Design der Formel für Betrag A entwickelt werden. Unter Betrag B sind vor allem noch Klärungen dazu nötig, welche Aktivitäten zu den Basismarketing– und Vertriebsaktivitäten zählen. Zudem muss der feste Renditeprozentsatz unter Betrag B definiert werden. Auch eine mögliche Safe-Harbour Rule unter Säule 1 steht im Raum, welche noch diskutiert werden muss.[254]

Innerhalb aller drei Beträge müssen effektive Massnahmen gegen Doppelbesteuerung sowie auch Massnahmen zur Streitvermeidung und Streitbeilegung erarbeitet werden. Diese sind zum momentanen Zeitpunkt noch nicht ausreichend definiert, sind laut OECD aber von grosser Bedeutung zur Erreichung eines Konsenses. Es muss zudem auch genau erörtert werden, wie die drei Beträge miteinander agieren, um eine Doppelbesteuerung ausschliessen zu können.[255] Weiters stellt die Administration und die Implementierung dieses Unified Approach die OECD vor weitere Herausforderungen. Es gilt zahlreiche Punkte zu klären, wie potentielle Änderungen der nationalen Gesetze, Änderungen der Steuerabkommen oder auch des automatischen Informationsaustausches.[256]

3.2.7. Ausblick

Es bleibt somit abzuwarten, wie sich der Unified Approach weiterentwickeln wird und ob ein globaler Konsens im Jahr 2020 erreicht werden kann. Der nächste Schritt der OECD zur Festigung des Unified Approach wird Anfang Juli 2020 in Berlin erwartet. Dort trifft sich der Inclusive Framework on BEPS zu einem Meeting, bei dem eine Einigung über die wesentlichen Kernpunkte erzielt werden soll. Ende Juli soll ein Treffen der G20-Finanzminister in Saudi-Arabien folgen, bei welchem eine Bestätigung sowie eine Anerkennung der Ergebnisse getroffen werden soll. Das Ziel der OECD ist

[246] Marcel Olbert und Christoph Sprengel, „Taxation in the Digital Economy: Recent Policy Developments and the Question of Value Creation: Working Paper," ZEW Discussion Paper, 19-010 (2019): 10, 23.

[247] Greil und Wargowske, „Pillar 1 of the Inclusive Framework's Work Programme: The Effect on the Taxation of the Digi-tal Economy and Reallocation of Taxing Rights," 506.

[248] Björn Heidecke et al., „Quo vadis Verrechnungspreise - die OECD auf Abwegen? Der Entwurf zur "Säule 1" v. 9.10.2019," IWB, Nr. 23 (2019): 940.

[249] Boehle und Scholtholt, „Gestaltungsoptionen zur gerechten Besteuerung von Digitalkonzernen," 923.

[250] Becker und Van der Ham, „Das neue (digitale) ABC der Verlagerung von Besteuerungsrechten in die Marktstaaten," 2545.

[251] Greil und Wargowske, „Pillar 1 of the Inclusive Framework's Work Programme: The Effect on the Taxation of the Digi-tal Economy and Reallocation of Taxing Rights," 513; Clemens Fuest, „Die G20-Beschlüsse: Durchbruch zu einer besseren internationalen Besteuerung?," DB, Nr. 29 (2019): M5.

[252] Schuster, „Der "Unified Approach" der OECD - ein Überblick und erste Einschätzungen," 519.

[253] Daurer, „Der Unified Approach der OECD," 557.

[254] OECD, „Outline of the Architecture of a Unified Approach on Pillar One," 22-23.

[255] Ebd., 23.

[256] Ebd., 24.

es, beim G20-Gipfel in Saudi-Arabien Ende November 2020 eine finale Entscheidung zu diesem Zwei-Säulen- Programm, und somit auch zum Unified Approach, zu treffen.[257] In einer Aussendung der OECD am 17.03.2020 wurde klargestellt, dass trotz der Coronavirus-Pandemie am Unified Approach weitergearbeitet wird und somit Anfang Juli 2020 in Berlin Entscheidungen über die Schlüsselfaktoren wie geplant getroffen werden sollen.[258] Da diese Aussendung der OECD am Anfang der Coronavirus-Pandemie veröffentlicht wurde, bleibt fraglich, inwieweit dieser Zeitplan eingehalten werden kann.

3.2.8. Kritik am Unified Approach

Für Schön nimmt diese Säule 1 der OECD ein Ausmass, welches weit über die Besteuerung der digitalen Wirtschaft hinausreicht, an. Er hebt aber hervor, dass dieser Ansatz sehr wohl attraktiv sein kann, da durch die Erfassung von Residualgewinnen die Steuerzahlungen vereinfacht und stabilisiert werden können. Zudem wird durch die Aufteilung des Residualgewinnes an die Marktjurisdiktionen der fiskalpolitische Bedarf der Marktstaaten gesättigt und die Marktmacht durch die Nachfrageseite berücksichtigt.[259] Eine andere Meinung vertritt dazu Ehlermann, der das Steuersubstrat auch als Zahlung für die im Staat mitgenutzte und benötigte Infrastruktur ansieht. Die Tatsache, dass diese Unternehmen steuerlich belangt werden, ohne jegliche Infrastruktur zu nutzen, scheint für Ehlermann bedenklich. Dabei verlieren die Staaten, in denen die Infrastruktur tatsächlich genutzt wird, an Besteuerungssubstrat und sind damit Verlierer der neuen Allokationsmethode.[260] Ditz und Pinkernell machen darauf aufmerksam, dass nicht klar scheint, warum anhand der Markt- und Konsumentennachfrage Besteuerungsrechte verteilt werden. Für die Autoren kann diese Verteilung auch anhand anderer Faktoren, die eine Erwirtschaftung von unternehmerischen Gewinnen ermöglichen, gemacht werden. Als Beispiel werden Produktions- und Kostenvorteile in Schwellen- und Entwicklungsländern sowie auch Forschung und Entwicklungstätigkeiten, die in kostengünstigen Ländern durchgeführt werden, genannt.[261]

Für Becker und Van der Ham hingegen sind Exportnationen wie Deutschland und Italien, die hohe Handelsüberschüsse verzeichnen, die grossen Verlierer unter Säule 1. Es bleibt offen, inwieweit Säule 2 in der Lage sein wird, dies zu kompensieren.[262] Auch Fuest hebt hervor, dass Exportnationen wie Deutschland in Zukunft einen grossen Teil der Gewinne in Marktländern wie China versteuern werden müssen. Folglich scheint es logisch, dass insbesondere Deutschland die Säule 2, eine effektive Mindestbesteuerung, umsetzen will, um so die Verluste einigermassen zu kompensieren.[263] Auch der Schweiz ist bewusst, dass den kleinen Nationen mit grosser Wertschöpfung das Besteuerungssubstrat entzogen wird. Die Auswirkungen auf die Steuereinnahmen der Schweiz lassen sich momentan noch nicht abschätzen, sie scheinen aber massiv zu sein. Berr, Bulard und Najafi sind zudem der Meinung, dass die Komplexität durch das ungewisse Zusammenspiel dieser verschiedenen Beträge weiter gesteigert wird.[264]

Für Schön wird durch den Unified Approach das Problem der Gewinnverlagerungen gelöst, da Unternehmen nicht mehr den Druck verspüren, Produktionsstätten in Niedrigsteuerländer zu verlagern.[265] Auch Fuest stimmt Schön in dieser Hinsicht zu, dass der Steuerwettbewerb nach unten mit der Einführung von Säule 1 teilweise beseitigt werden kann, da sich Umsatzmärkte nicht so einfach verschieben lassen wie beispielsweise Patente oder Produktionsstätten.[266]

Daurer macht darauf aufmerksam, dass, obwohl eine Einigung der technischen Aspekte erzielt werden kann, immer noch das Problem der Umsetzung bleibt. Für die Autorin scheint die Umsetzung anhand eines „MLI 2“, also anhand eines multilateralen Instrumentes, plausibel. Hierbei entsteht jedoch das zusätzliche Problem, dass nicht alle Mitgliedstaaten des Inclusive Framework DBA untereinander abgeschlossen haben. Für Daurer kann die Wirksamkeit des Unified Approach jedoch nur global verwirklicht werden, weshalb dies zu einem globalen multilateralen Abkommen führen würde, was jedoch komplettes Neuland innerhalb des Steuerrechts bedeutet. Daraus ergibt sich auch zudem die Frage, wie das nationale Recht ausgestaltet wird. Denn DBA entfalten lediglich Schrankenwirkung, was bedeutet, dass diese nur einschränken können, wenn bereits ein nationales Recht besteht.[267] Berr, Bulard und Najafi weisen zudem darauf hin, dass in der Praxis die Methoden zur Vermeidung der Doppelbesteuerung zu Problemen führen können, wenn beispielsweise der neue Nexus die Besteuerung eines MNE begründet, jedoch mangels einer physischen Präsenz dieses MNE das DBA keine Wirkung entfalten könnte.[268]

Für Ehlermann ist hingegen das Bedürfnis der dringenden Lösungssuche nur begrenzt verständlich, da sich die meisten BEPS-Massnahmen noch in Umsetzung befinden

[257] Ebd.; Wenz, „The Future of International Taxation,“ 42.

[258] OECD, „Coronavirus (Covid-19): Update on OECD tax work,“ Zugriff am 31.03.2020, http://www.oecd.org/tax/beps/coronavirus-covid-19-update-on-oecd-tax-work.htm.

[259] Schön, „Seminar F (IFA/OECD): Internationale Besteuerung der digitalisierten Wirtschaft,“ 650.

[260] Ehlermann, „Welche Richtung nimmt die Besteuerung digitaler Geschäftsmodelle?,“ M5.

[261] Xaver Ditz und Reimar Pinkernell, „Neudefinition internationaler Besteuerungsrechte durch das OECD Inclusive Framework on BEPS - Eine Würdigung aus deutscher Sicht,“ ISR, Nr. 11 (2019): 382.

[262] Becker und Van der Ham, „Das neue (digitale) ABC der Verlagerung von Besteuerungsrechten in die Marktstaaten,“ 2544.

[263] Fuest, „Die G20-Beschlüsse: Durchbruch zu einer besseren internationalen Besteuerung?,“ M4.

[264] Fabian Berr, Céline Bulard und Niloufar Najafi, „OECD-Vorschlag für eine einheitliche Besteuerung der digitalisierten Wirtschaft,“ Expert Focus, Nr. 3 (2020): 153.

[265] Schön, „Seminar F (IFA/OECD): Internationale Besteuerung der digitalisierten Wirtschaft,“ 650.

[266] Fuest, „Die G20-Beschlüsse: Durchbruch zu einer besseren internationalen Besteuerung?,“ M4.

[267] Daurer, „Der Unified Approach der OECD,“ 556-557.

[268] Berr, Bulard und Najafi, „OECD-Vorschlag für eine einheitliche Besteuerung der digitalisierten Wirtschaft,“ 153.

oder gerade erst umgesetzt werden. Hierbei sollte man zuerst die Wirkungsweise dieser Massnahmen abwarten, bevor krampfhaft nach einer neuen Lösung gesucht wird.[269]

Laut einer Studie der Universität Mannheim und des Leibniz-Zentrums für Europäische Wirtschaftsforschung Mannheim stellt die Einführung einer Digitalsteuer eine wesentliche Gefährdung der Wettbewerbsfähigkeit und der Rentabilität digitaler Unternehmen dar. Aufgrund des Vorschlags der EU zur Implementierung einer Digitalsteuer reagierten Kapitalmärkte in einem nicht absehbaren Ausmass und Marktwerte vieler digitaler Unternehmen fielen rasant. Es muss somit beachtet werden, dass Digitalsteuern keine negativen Folgen für Unternehmen haben und die Wirkung im Vorhinein detailliert untersucht wird. Es wird somit zur Vorsicht bei der Einführung einer Digitalsteuer geraten, da sich dies zudem auch in einer geringeren Investitionsbereitschaft und geringeren Wachstumsmöglichkeiten niederschlagen kann.[270] Auch Ehlermann macht darauf aufmerksam, dass sich das Steuerrecht zum Investitionshemmnis entwickeln kann, wenn nicht schnellstmöglich klare Regeln innerhalb eines internationalen Konsenses geschaffen werden.[271]

Diese Diskussion um Besteuerungsrechte stellt ein hoch brisantes politisches Thema dar und kann nur durch Kompromisse gelöst werden. Die Meinungen bezüglich des internationalen Steuerrechts sind laut Kreienbaum höchst verschiedene und bringen die „Vielzahl unterschiedlicher standortökonomischer und fiskalischer Perspektiven“ zum Ausdruck. Während manche Staaten den Status quo als ausreichend empfinden, kritisieren andere die zunehmende Unbrauchbarkeit des Fremdvergleichsgrundsatzes. Zudem gibt es Staaten, welche Säule 1 unterstützen und sich vollkommen gegen Säule 2 aussprechen oder umgekehrt. Für Kreienbaum scheinen die zukünftigen Entwicklungen dieses OECD-Projektes von der Kompromissfähigkeit der Staaten abhängig.[272] Auch für Fuest kann ein Scheitern des Projektes nicht ausgeschlossen werden, da fraglich ist, ob die Gewinner und Verlierer dieses Ansatzes die Umverteilung des Besteuerungssubstrats auch akzeptieren.[273]

4. Auswirkungen des Unified Approach auf die Besteuerung - Fallbetrachtung anhand eines fiktiven Geschäftsmodelles

Im nachfolgenden Kapitel wird nun die fiktive Unternehmensgruppe X, deren Unternehmenstätigkeit auf einem digitalen Geschäftsmodell basiert, als Grundlage zur Bewertung des Unified Approach herangezogen. Mithilfe dieser Fallbetrachtung wird geprüft, ob der Vorschlag der OECD in der Lage ist, die vorhandenen Probleme und Herausforderungen bei der Besteuerung digitaler Geschäftsmodelle, die anfangs in dieser Arbeit identifiziert wurden, zu lösen bzw. zu verringern. Die Unternehmensgruppe X soll sowohl das Problem der Vereinfachung von Gewinnverlagerungs- und Gewinnvermeidungsstrategien (BEPS) (siehe 2.3.1), das Problem des Konzeptes der physischen Präsenz (2.3.2), das Problem des komplexen Wertschöpfungsbeitrages innerhalb der digitalen Wirtschaft (siehe 2.3.3) sowie auch das Problem der Qualifizierung von Einkünften (siehe 2.3.4) darstellen. Das Fallbeispiel wird sehr vereinfacht dargestellt, um den Fokus auf die relevanten und neuen Aspekte des Unified Approach zu lenken.

Zudem folgt unter Kapitel 4.2 eine Abwandlung des Fallbeispieles aus Kapitel 4.1. Die Unternehmenstätigkeit der Unternehmensgruppe Y basiert in diesem Beispiel 2 nicht mehr auf einem digitalen Geschäftsmodell, sondern auf einem traditionellen Geschäftsmodell. Das Unternehmen stellt somit ein Unternehmen der „Old Economy“ dar. In diesem abgewandelten Beispiel wird insbesondere auf die Nexus-Thematik näher eingegangen. Da die Berechnung der Beträge A, B und C zu keinen grossen Unterschieden im Hinblick auf das erste Beispiel führen würde, wird diese vernachlässigt.

Der Routineanteil in Prozent bzw. die Routinegewinnmarge bei der Berechnung des Betrages A wird in diesem Beispiel mit 10 % angenommen. Der Anteil des Residualgewinnes, welcher schlussendlich auf die Marktstaaten verteilt wird, wird mit 20 % angenommen. Diese Annahmen werden seitens der OECD in der vorläufigen Analyse und Folgenabschätzung zu Säule 1 verwendet, weshalb auch in dieser Fallbearbeitung darauf zurückgegriffen wird.[274]

4.1. Beispiel 1: Das digitale Geschäftsmodell der Unternehmensgruppe X

Die Unternehmensgruppe bzw. der Konzern X ist ein multinationaler digitaler Konzern[275], welcher in Abbildung 14 näher dargestellt wird. Das Mutter-Unternehmen B-Ltd. verfügt über einen Sitz auf den Bermudas (Niedrigsteuerland), wobei sich dort auch der Ort der Geschäftsleitung befindet und alle strategischen Entscheidungen werden auf den karibischen Inseln getroffen. Das Geschäftsmodell besteht darin, Cloud-Computing Dienstleistungen anzubieten. Dies bedeutet, dass die Unternehmensgruppe X ihren Kunden über einen Server Speicherkapazitäten und weitere digitale Dienstleistungen auf Abruf zur Verfügung stellt. Die Immaterialgüterrechte, die für den Cloud-Computing-Service benötigt werden, werden von der B-Ltd. gehalten. Die B-Ltd. überlässt der NL-B.V. Lizenzrechte für das Cloud-Geschäft. Weiters wird die

[269] Ehlermann, „Welche Richtung nimmt die Besteuerung digitaler Geschäftsmodelle?,“ M5.

[270] Christopher Ludwig, Christoph Sprengel und Daniel Klein, „Eine Digitalsteuer gefährdet Wettbewerbsfähigkeit und Rentabilität,“ ZEW News, Nr. 12 (2019): 7.

[271] Ehlermann, „Welche Richtung nimmt die Besteuerung digitaler Geschäftsmodelle?,“ M5.

[272] Martin Kreienbaum, „Fortschritte bei der Digitalbesteuerung - Zweisäulenstrategie in der Diskussion,“ IStR, Nr. 4 (2019): 123.

[273] Fuest, „Die G20-Beschlüsse: Durchbruch zu einer besseren internationalen Besteuerung?,“ M5.

[274] OECD, „Tax Challenges arising from the Digitalisation of the Economy,“ 12.

[275] In Anlehnung an OECD, „Secretariat Proposal for a “Unified Approach” under Pillar One,“ 11-12.

ganze Programmiertätigkeit für das Cloud-Geschäft im Ansässigkeitsstaat des Mutterunternehmens, also auf den Bermudas, durchgeführt.

Die B-Ltd. besitzt zu 100 % eine Beteiligung an der NL-B.V. in den Niederlanden, die dort ihren Sitz und Ort der gewöhnlichen Geschäftstätigkeit aufweist. Die Tochtergesellschaft NL-B.V. „liefert" diese Cloud-Services an verschiedene Marktjurisdiktionen innerhalb und ausserhalb Europas. Die niederländische Tochter ist für das Marketing und die Distribution der Unternehmensgruppe X zuständig und verkauft die Dienstleistung in den Niederlanden wie auch per Fernabsatz in Österreich, Deutschland und den USA. Weiters muss die niederländische Tochtergesellschaft annahmegemäss fremdübliche Lizenzzahlungen in Höhe von 950 Mio. EUR an die Muttergesellschaft B-Ltd. leisten.

Es wird davon ausgegangen, dass die Unternehmensgruppe X einen konsolidierten Unternehmensumsatz von 1.000 Mio. EUR weltweit erzielt. Annahmegemäss wird ein Unternehmensgewinn von 250 Mio. EUR erwirtschaftet und somit eine Gewinnmarge von 25 % verzeichnet.

Die Umsätze des operativen Geschäftes, die die Konzerngesellschaft X in den verschiedenen Ländern generiert, können aus Tabelle 2 entnommen werden:

Das Beispiel der Unternehmensgruppe X soll einerseits die Problematik rund um BEPS, also die Möglichkeit der Gewinnverlagerungsmassnahmen in Niedrigsteuerländer (Bermudas) sowie das Ausnutzen von Steuersatzarbitrage zwischen einem Hoch- und Niedrigsteuerland verdeutlichen. Andererseits soll aber auch die Problematik rund um das Konzept der physischen Präsenz beschrieben werden, da die Unternehmensgruppe X „remote" an verschiedene Staaten (Deutschland, Österreich und USA) verkauft. Zudem soll auch das Problem des komplexen Wertschöpfungsbeitrages innerhalb digitaler Geschäftsmodelle dargestellt werden. Bei vielen digitalen Geschäftsmodellen stellt sich immer wieder die Frage, ob und wie der Ort der Wertschöpfung mit dem Ort der Besteuerung zusammengebracht werden kann bzw. soll. Ausserdem kann dieses Beispiel auch das Problem der Qualifizierung von Einkünften hinsichtlich Cloud-Computing aufzeigen.

Die folgende Tabelle 3 veranschaulicht die Annahmen über die Körperschaftsteuersätze der verschiedenen Länder, welche in der untenstehenden Fallbearbeitung genutzt werden.

4.1.1. Besteuerung nach momentanem Recht

Da die Besteuerung nach momentanem Recht am Konzept der physischen Präsenz anknüpft, können zum jetzigen Zeitpunkt lediglich auf den Bermudas und in den Niederlanden die Gewinne versteuert werden. Die Marktstaaten, also die Ansässigkeitsstaaten der Nutzer dieser Clouddienste, haben keinerlei Recht auf eine Besteuerung von Unternehmensgewinnen, da der physische Anknüpfungspunkt fehlt und somit kein Nexus vorhanden ist. Dies bedeutet in Bezug auf das oben genannte Beispiel, dass Deutschland, Österreich und die USA kein Besteuerungsrecht haben.

Die Tochtergesellschaft NL-B.V. ist aufgrund des Sitzes in den Niederlanden unbeschränkt steuerpflichtig in den Niederlanden mit ihren gesamten Unternehmensgewinnen. Die Umsätze der operativen Tätigkeit, also die Umsätze aus den Niederlanden, den USA, Deutschland und Österreich, werden somit der niederländischen Gesellschaft zugerechnet. Somit wäre der Umsatz der Tochtergesellschaft 1.000 Mio. EUR. Da die niederländische Tochter jedoch Lizenzzahlungen in Höhe von 950 Mio. EUR an die Mutter auf den Bermudas zahlen muss, wird der Unternehmensgewinn der NL-B.V. auf 50 Mio. EUR reduziert. Nach niederländischem Recht gibt es eine IP-Box Regelung von 20 %, sodass von diesen 50 Mio. EUR Gewinn lediglich 20 % (10 Mio. EUR) besteuert werden können.

Die B-Ltd. ist mit ihrem Unternehmensgewinn auf den Bermudas unbeschränkt steuerpflichtig. Dieser Unternehmensgewinn beinhaltet die Lizenzeinnahmen der niederländischen Tochter (950 Mio. EUR) und kann somit vollumfänglich auf den Bermudas besteuert werden. Der Unternehmensgewinn wird um allfällige Aufwendungen verringert (Annahme: 750 Mio. EUR), sodass ein Unternehmensgewinn der B-Ltd. von 200 Mio. EUR entsteht. Somit ist der konsolidierte Konzerngewinn des Konzernes X 250 Mio. EUR (50 Mio. EUR Niederlande und 200 Mio. EUR Bermudas). Die B-Ltd. ist zudem nicht beschränkt steuerpflichtig in den Niederlanden, da nach niederländischem Recht keine Quellensteuer auf Lizenzzahlungen ins Ausland erhoben werden, weshalb die Lizenzzahlungen der NL-B.V. unversteuert auf die Bermudas fliessen.

Da die Bermudas als Niedrigsteuerland keine Körperschaftsteuer auf Unternehmensgewinne erheben, bleiben die 200 Mio. EUR der B-Ltd. unversteuert. Der Unternehmensgewinn der NL-B.V. von 50 Mio. EUR wird in den Niederlanden aufgrund einer IP-Box Regelung lediglich zu 20 % besteuert. Bei Annahme eines Körperschaftsteuersatzes von 12,5 % und einer gesenkten Bemessungsgrundlage von 10 Mio. EUR (20 % von 50 Mio. EUR) ergibt dies eine Steuerbelastung der NL-B.V. in den Niederlanden von 1,25 Mio. EUR. Dies ist somit auch die gesamte Steuerbelastung des Konzernes X. Die daraus resultierende Konzernsteuerquote der Unternehmensgruppe X beträgt 0,5 %.

Gesamtsteuerbelastung: 1,25 Mio. EUR
Konzernsteuerquote: 0,5 %

4.1.2. Mögliche Besteuerung nach dem Unified Approach

In diesem Beispiel wird davon ausgegangen, dass die Unternehmensgruppe X, als Anbieter eines Automated digital service und mit einem Konzernumsatz von 1.000 Mio. EUR, in den Anwendungsbereich des Betrages A fällt. Es wird angenommen, dass alle relevanten Schwellenwerte überschritten worden sind. Es wird auch davon ausgegangen, dass in allen Marktstaaten ein Nexus nach Betrag A begründet wird. In weiterer Folge stellt sich die Frage, wie hoch die Beträge A, B und C sind und wie diese an die Marktjurisdiktionen aufgeteilt werden können.

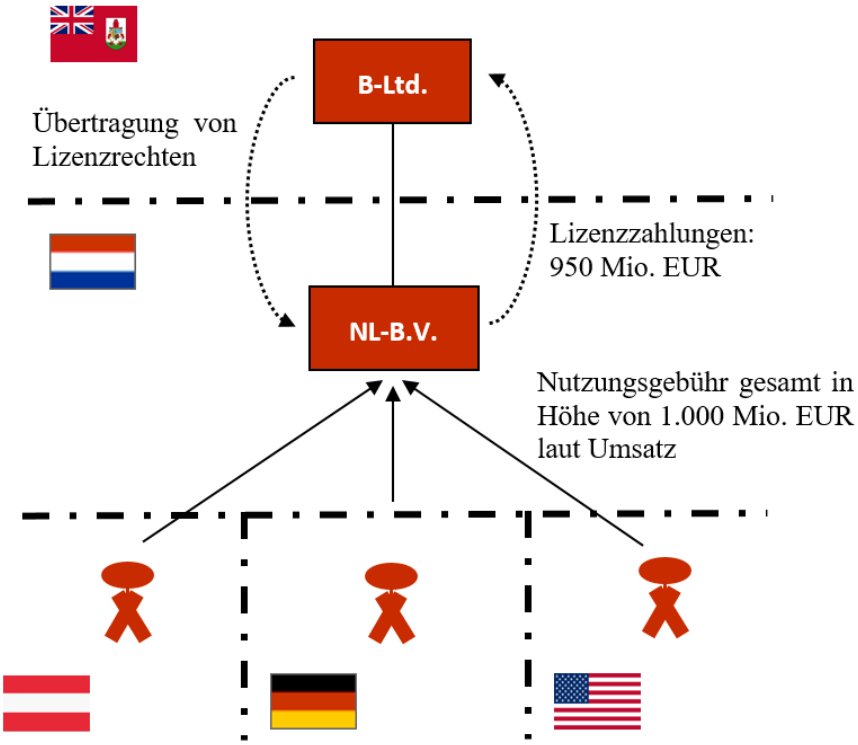


Abbildung 14: Unternehmensgruppe X
Eigene Abbildung

Tabelle 2: Umsatz Unternehmensgruppe X

Gesamtumsatz der Unternehmensgruppe X		
Umsatz Bermudas	0 Mio. EUR	0 %
Umsatz Niederlande	200 Mio. EUR	20 %
Umsatz Deutschland	300 Mio. EUR	30 %
Umsatz Österreich	150 Mio. EUR	15 %
Umsatz USA	350 Mio. EUR	35 %
Gesamt	1.000 Mio. EUR	100 %

Tabelle 3: Annahme Körperschaftsteuersätze

Bermudas	0 %
Deutschland	15 %
Niederlande	12,5 %
Österreich	25 %
USA	20 %

Es werden nun zuerst die Beträge B und C berechnet, da auch in der Praxis zuerst diese Beträge anhand den Verrechnungspreisregeln zugeteilt werden. Anschliessend wird der Betrag A berechnet und auf die verschiedenen Marktstaaten verteilt. Die Vorgehensweise der Berechnungen sowie die Ausgestaltung der Beträge A, B und C wird in Anlehnung an Greil und Wargowske[276] durchgeführt. Für die Beträge B und C sind keine Änderungen von der Besteuerung nach momentanem Recht und der Besteuerung nach dem Unified Approach ersichtlich, da diese weiterhin auf den bestehenden Verrechnungspreisregeln basieren.

[276] Greil und Wargowske, „Pillar 1 of the Inclusive Framework's Work Programme: The Effect on the Taxation of the Digi-tal Economy and Reallocation of Taxing Rights," 508-510.

Betrag B

Der Betrag B ist weiterhin von dem Konzept der physischen Präsenz abhängig. Im Konzern X kann somit bezüglich Betrag B lediglich die B-Ltd. auf den Bermudas bzw. die NL-B.V. in den Niederlanden steuerpflichtig werden. Aufgrund des fehlenden steuerlichen Anknüpfungspunktes der Unternehmensgruppe X in Deutschland, Österreich und den USA kann dort kein Betrag B erhoben werden.

Die OECD hat hierbei noch nicht klar festgestellt, welche Aktivitäten in die Basismarketing- und Vertriebsaktivitäten fallen. Es wird in diesem Beispiel die Annahme getroffen, dass nur die NL-B.V. diese relevanten Aktivitäten unter Betrag B durchführt und somit auch in den Niederlanden mit dem Betrag B steuerpflichtig ist. Die Bemessungsgrundlage für diese Aktivitäten unter Betrag B beläuft sich annahmegemäss auf 10 Mio. EUR, welche in Tabelle 4 ersichtlich ist. Diese 10 Mio. EUR werden mit dem nationalen Körperschaftsteuersatz der Niederlande besteuert.

Seitens der OECD ist bis zum jetzigen Zeitpunkt noch nicht näher geklärt, wie diese Bemessungsgrundlage für Betrag B bzw. die Routinevergütung unter Betrag B berechnet wird. PwC gibt hierbei in ihrem Input-Statement zum Public Consultation Document einige Vorschläge, wie die Bemessungsgrundlage für Betrag B quantifiziert werden könnte. Beispielsweise könnte hierbei ein gewisser Prozentsatz des Return-on-Sales (ROS) oder auch ein gewisser Prozentsatz der Investitionen, die ein MNE tätigen muss, um Zugang zu einem gewissen Markt zu erhalten, als Bemessungsgrundlage für Betrag B herangezogen werden.[277]

Betrag C

Da auch der Betrag C nur aufgrund eines physischen Anknüpfungspunktes steuerlich geltend wird, kann der Betrag C somit nicht von Deutschland, Österreich oder den USA gefordert werden. Es besteht aber die Möglichkeit, dass die Niederlande die Routinevergütung aus Betrag B als zu niedrig empfindet und somit auf eine höhere Entschädigung plädiert. Dies kann in diesem Fallbeispiel von den Niederlanden gefordert werden, da annahmegemäss aufgrund übermässiger Marketingaktivitäten ein Markenwert für die Unternehmensgruppe X aufgebaut werden konnte.

Es wird in diesem Beispiel davon ausgegangen, dass laut dem Fremdvergleichsgrundsatz 40 Mio. EUR als Bemessungsgrundlage für Betrag C anfallen. Somit liegt die Bemessungsgrundlage des Betrages C bei 40 Mio. EUR in den Niederlanden, wie aus Tabelle 5 ersichtlich wird.

Die gesamte Bemessungsgrundlage für Betrag B und C liegt bei 50 Mio. EUR in den Niederlanden und bleibt somit gleich hoch wie die Bemessungsgrundlage nach der Besteuerung nach momentanem Recht. Dies resultiert aus der Tatsache, dass die Verrechnungspreisrichtlinien weiterhin wie bisher genutzt werden, da diese nicht angepasst oder aktualisiert werden. Die Beträge B und C führen somit zu keinen Veränderungen in der Gewinnverteilung, da auch anhand bestehender Verrechnungspreisrichtlinien die Besteuerung dieser Vergütungen unter Betrag B und C bei den Niederlanden liegt.

Betrag A

1. Ermittlung des Residualgewinnes
   Zunächst muss der Residualgewinn der Unternehmensgruppe X bestimmt werden. Der Unternehmensumsatz wurde anhand eines konsolidierten Konzernabschlusses berechnet und beträgt 1.000 Mio. EUR. Der Gesamtgewinn des Konzernes X entspricht dem Gewinn vor Steuern, der als Grundlage für Betrag A herangezogen wird und beträgt annahmegemäss 250 Mio. EUR. Die Ermittlung des Residualgewinnes ist in Tabelle 6 bzw. Tabelle 7 abgebildet.
   15 % von 1.000 Mio. EUR = 150 Mio. EUR Residualgewinn
2. Zuordnung des Residualgewinnes
   Seitens der OECD ist noch nicht klar, welcher Anteil des Residualgewinnes an die Marktstaaten verteilt wird. Hierbei wird unterschieden, welcher Teil des Residualgewinnes auf die Wertschöpfung zurückzuführen ist und welcher Teil auf andere Faktoren, wie Kapital und Risiko, entfällt. Für diese Fallbetrachtung wird ein Prozentsatz von 80 %, welcher nicht umsatz-orientiert ist, angenommen. Dies bedeutet, dass 20 % von 150 Mio. EUR (30 Mio. EUR) an die Marktjurisdiktionen verteilt werden.

$$30 \text{ Mio. EUR} + 120 \text{ Mio. EUR} = 150 \text{ Mio. EUR} \quad (3)$$

   30 Mio. EUR - Anteil der Marktjurisdiktion am Residualgewinn
   120 Mio. EUR - nicht-umsatzorientierter Anteil am Residualgewinn
   150 Mio. EUR - zu verteilender Residualgewinn aus Nicht-Routinetätigkeiten
3. Verteilung des Residualgewinnes Nun muss der Residualgewinn von 30 Mio. EUR anhand eines Verteilungsschlüssels, basierend auf Umsätzen, auf die Marktstaaten verteilt werden. Dies wird in Tabelle 8 näher verdeutlicht.

Steuerbelastung anhand des Unified Approach

Während die Beträge B und C weiterhin an den Verrechnungspreisen anknüpfen, stellt Betrag A jedoch eine Überlagerung des Systems dar und wird „on top“ von den Marktstaaten versteuert. Der Residualgewinn, der an die Marktstaaten verteilt wird, belastet das MNE somit zusätzlich. Die Muttergesellschaft B-Ltd. wird weiterhin, wie auch nach momentanem Recht, mit 200 Mio. EUR auf den Bermudas besteuert. Aufgrund des Körperschaftsteuersatzes von 0 % bleibt dieser Gewinn unversteuert. Innerhalb dieses Beispieles kommt es somit zu keiner Doppelbesteuerung aufgrund des Betrages A, da die Bermudas eine Körperschaftsteuer von 0 % aufweisen. Jedoch kann der Betrag A sehr wohl zu

[277] PricewaterhouseCoopers, „Reaction to the OECD secretariat's consultation paper on the unified approach under Pillar 1,“ 26-29.

Tabelle 4: Bemessungsgrundlage Betrag B – Unternehmensgruppe X

	Bemessungsgrundlage Betrag B
Niederlande	10 Mio. EUR
	10 Mio. EUR

Tabelle 5: Bemessungsgrundlage Betrag C – Unternehmensgruppe X

	Bemessungsgrundlage Betrag C
Niederlande	40 Mio. EUR
	40 Mio. EUR

einer Doppelbesteuerung führen, falls die Muttergesellschaft in einem Hochsteuerland ansässig ist und mit dem Welteinkommensprinzip besteuert wird. Hierbei würden die Gewinne sowohl im Ansässigkeitsstaat der Muttergesellschaft als auch in den Marktstaaten besteuert werden.[278]

Es ist noch nicht geklärt, wie das Problem der Doppelbesteuerung unter Betrag A in Zukunft gelöst werden kann. Die Lösung des Problemes der Doppelbesteuerung wird auch von der OECD mit höchster Priorität eingestuft. Die OECD[279] diskutiert beispielsweise ein Senken der Bemessungsgrundlage um den Gewinn, der in den Marktstaaten versteuert wird bzw. eine Anrechnung der Steuer, die in den Marktstaaten gezahlt wird. Würde man also davon ausgehen, dass die B-Ltd. als Muttergesellschaft steuerpflichtig für den Betrag A wäre, könnte eine mögliche Lösung der Doppelbesteuerung unter Betrag A das Senken der Bemessungsgrundlage um 30 Mio. EUR auf den Bermudas sein. Dies würde in diesem Beispiel jedoch zu keinerlei Änderungen führen, da die Bermudas einen Körperschaftsteuersatz von 0 % aufweisen. Aus diesem Grund wird auf die Berichtigung der Bemessungsgrundlage der B-Ltd. in diesem Beispiel verzichtet.

Die Steuerbemessungsgrundlagen der Beträge A, B und C können der Tabelle 9 entnommen werden.

Nun werden die Bemessungsgrundlagen in den jeweiligen Ländern mit den nationalen Körperschaftsteuersätzen dieser Länder besteuert. Die Bemessungsgrundlage der Niederlande in der Höhe von 56 Mio. EUR wird, wie bei der Besteuerung nach momentanem Recht, aufgrund einer IP-Box Regelung nur im Umfang von 20 % (11,2 Mio. EUR) besteuert.

Daraus ergibt sich eine Gesamtsteuerbelastung der Unternehmensgruppe X von 5,975 Mio. EUR, wie in Tabelle 10 dargestellt ist. Die Konzernsteuerquote liegt mit der Besteuerung anhand des Unified Approach bei 2,39 %.

Gesamtsteuerbelastung: 5,975 Mio. EUR
Konzernsteuerquote: 2,39 %

[278] Vgl. Ebd., 22.

[279] OECD, „Outline of the Architecture of a Unified Approach on Pillar One," 14.

4.1.3. Bewertung des Unified Approach anhand der Unternehmensgruppe X

Die erste genannte Herausforderung bei der Besteuerung der Digital Economy ist das Problem der vereinfachten Gewinnverlagerungs- und Gewinnverschiebungsmassnahmen und den daraus entstandenen niedrigeren Steuerlasten für digitale und multinationale Konzerne. Bezüglich der Unternehmensgruppe X kann gesagt werden, dass die BEPS-Probleme durch diesen Unified Approach mit hoher Wahrscheinlichkeit abnehmen werden. Wie bereits durch Fuest[280] erwähnt, lassen sich User, Marktjurisdiktionen und Umsatzmärkte nicht so leicht verschieben wie Betriebsstätten. Hierbei ist es eindeutig, dass eine Besteuerung in Steueroasen zumindest in Teilen vermieden werden kann. Wie das Beispiel 1 jedoch auch zeigt, wird je nach der Entscheidung der OECD, wie hoch die festgelegten Prozentsätze sind, nur ein kleiner Teil des Gesamtgewinnes des Konzernes auf die Marktstaaten aufgeteilt. Zudem kommt auch das Problem hinzu, dass die den Marktjurisdiktionen zugeteilten Gewinne mit dem nationalen Körperschaftsteuersatz besteuert werden. Es kann also trotzdem der Fall sein, dass Staaten, die einen Betrag A zugewiesen bekommen würden, diesen dann nicht oder nur niedrig besteuern. Weiters wurde von Greil und Wargowske[281] bereits erwähnt, dass ein Steuersystem in diesem Komplexitätsgrad neue Schlupflöcher bezüglich BEPS schaffen kann. Somit bleibt es abzuwarten, ob sich eine Vielzahl der BEPS-Probleme durch diesen Ansatz lösen oder auch neue BEPS-Probleme hinzugewonnen werden.

Des Weiteren wird sich die Konzernsteuerquote und somit die Steuerbelastung durch die Besteuerung anhand des Unified Approach erhöhen, wie bei der Unternehmensgruppe X ersichtlich ist. Hierbei steigt die Konzernsteuerquote von 0,5 % auf 2,39 %, wie in Tabelle 11 ersichtlich.

Dies wurde auch seitens der OECD[282] bei ihrem Update zur wirtschaftlichen Analyse und Folgenabschätzung des

[280] Fuest, „Die G20-Beschlüsse: Durchbruch zu einer besseren internationalen Besteuerung?," M4.

[281] Greil und Wargowske, „Pillar 1 of the Inclusive Framework's Work Programme: The Effect on the Taxation of the Digi-tal Economy and Reallocation of Taxing Rights," 513.

[282] OECD, „Tax Challenges arising from the Digitalisation of the Economy," 13.

Tabelle 6: Ermittlung Residualgewinn 1 – Unternehmensgruppe X

Rechenweg 1:	
Gewinn vor Steuern Konzern X	250 Mio. EUR
– 10 % Routinegewinnmarge (1.000 Mio. EUR * 10 %)	-100 Mio. EUR
= Residualgewinn	150 Mio. EUR

Tabelle 7: Ermittlung Residualgewinn 2 – Unternehmensgruppe X

Rechenweg 2:	
Gewinnspanne (Gewinn : Umsatz) in Prozent	25 %
– Routineanteil in Prozent (Annahme)	-10 %
= zu verteilender Residualgewinn aus Nicht-Routinetätigkeiten in Prozent	15 %

Tabelle 8: Verteilung Residualgewinn – Unternehmensgruppe X

	Anteil	Bemessungsgrundlage Betrag A
Residualgewinn gesamt	100 %	30 Mio. EUR
Anteilig Niederlande	20 %	6 Mio. EUR
Anteilig Deutschland	30 %	9 Mio. EUR
Anteilig Österreich	15 %	4,5 Mio. EUR
Anteilig USA	35 %	10,5 Mio. EUR

Tabelle 9: Bemessungsgrundlagen ABC-Beträge – Unternehmensgruppe X

Bemessungsgrundlagen der Unternehmensgruppe X nach ABC-Beträgen				
	Betrag A	Betrag B	Betrag C	Summe
Bermudas				0 Mio. EUR
Niederlande	6 Mio. EUR	40 Mio. EUR	10 Mio. EUR	56 Mio. EUR
Deutschland	9 Mio. EUR			9 Mio. EUR
Österreich	4,5 Mio. EUR			4,5 Mio. EUR
USA	10,5 Mio. EUR			10,5 Mio EUR
Summe	30 Mio. EUR	40 Mio. EUR	10 Mio. EUR	80 Mio. EUR

Tabelle 10: Bemessungsgrundlage gesamt – Unternehmensgruppe X

	Bemessungsgrundlage gesamt	Steuersatz	Steuerbelastung
Bermudas	200 Mio. EUR	0 %	0 Mio. EUR
Niederlande	11,2 Mio. EUR	12,5 %	1,4 Mio. EUR
Deutschland	9 Mio. EUR	15 %	1,35 Mio EUR
Österreich	4,5 Mio EUR	25 %	1,125 Mio EUR
USA	10,5 Mio EUR	20 %	2,1 Mio. EUR
Summe			5,975 Mio. EUR

Tabelle 11: Steuerbelastung, Vergleich momentanes Recht – Unified Approach

	Momentanes Recht	Unified Approach
Steuerbelastung	1,25 Mio. EUR	5,975 Mio. EUR
Konzernsteuerquote	0,5 %	2,39 %

Zwei-Säulen-Ansatzes berechnet. Die geringe Steuerlast der digitalen Welt führte immer wieder zu zahlreichen Diskussionen. Der Unified Approach wird hierbei Abhilfe schaffen und die Steuerlast der digitalisierten Welt zumindest etwas anheben. Der Vollständigkeit halber muss auch erwähnt werden, dass dieses Problem der zu geringen Steuerlast der MNE vor allem durch Säule 2 des Zwei-Säulen-Ansatzes näher verfolgt wird. Säule 2 wird laut OECD zu einem wesentlichen und auch grösseren Anstieg der Steuerquote als Säule 1 führen.[283]

Das zweite genannte Problem, die Konzentration des internationalen Steuerrechts auf das Konzept der physischen Präsenz, kann durch diesen Ansatz gelöst und überwunden werden. Die Marktjurisdiktionen haben nun ein Besteuerungsrecht und einen steuerrechtlichen Anknüpfungspunkt, ohne eine physische Präsenz in einem Staat aufzuweisen. Es muss aber erwähnt werden, dass, wie in Beispiel 1 verdeutlicht, lediglich ein kleiner Teil des Gewinnes in die Marktstaaten verteilt wird und dieses Besteuerungsrecht der Marktstaaten dementsprechend klein sein kann. Jedoch wird durch den Unified Approach die internationale Unternehmensbesteuerung wesentlich verändert und kann nun auch ohne das Konzept der physischen Präsenz Besteuerungsrechte garantieren.

Das Problem des komplexen Wertschöpfungsbeitrages kann durch den Unified Approach nur teilweise gelöst werden, so wie das Beispiel der Unternehmensgruppe X verdeutlicht. Hierbei bleibt weiter die elementare Frage, wie Wertschöpfung definiert werden kann und welche Aktivitäten zur Wertschöpfung beitragen. Innerhalb des Beispieles 1 scheint es fraglich, warum hier ein Teil des Gewinnes den Marktstaaten zugeteilt wird, da die relevante Wertschöpfung der Unternehmensgruppe X, die Programmierarbeit für den Cloud-Computing Dienst, am Unternehmenssitz auf den Bermudas getätigt wird. Es bleibt also zu diskutieren, ob und wie die User in dem Marktstaat eine Wertschöpfung generieren und ob somit die Besteuerung eines Teiles des Gewinnes im Markstaat erforderlich ist.

Die Wertschöpfung ist in allen digitalen Geschäftsmodellen sehr unterschiedlich, da auch die User einen unterschiedlichen Anteil dazu beitragen. Auch beim Cloud-Computing Dienst der Unternehmens-gruppe X lassen sich unterschiedliche Standpunkte einnehmen. Einerseits findet die relevante Programmiertätigkeit für das Geschäftsmodell nur auf den Bermudas statt. Andererseits tragen die Nutzer auch indirekt zur Wertschöpfung bei, da insbesondere durch eine immer grösser werdende Anzahl von Benutzern das Cloud-Computing erst effektiv und effizient wird, da sich damit die Kosten für das einzelne Unternehmen reduzieren.

Hinsichtlich Beispiel 1 wird deutlich, dass der Ort der Wertschöpfung und der Ort der Besteueurung nach Bauer et al.[284] nicht auseinanderfallen würde, da keine Monetarisierung der Daten bzw. auch keine Datenerhebung, sondern lediglich Konsum der Cloud-Computing Dienstleistung, stattfindet. Somit wäre kein steuerrechtliches Problem vorhanden und das alte Besteuerungsrecht würde dieses Geschäftsmodell richtig und auch ausreichend erfassen. Ditz und Pinkernell[285] vertreten die Meinung, dass der Wertschöpfungsbeitrag der User erst dann besteht, wenn eine Monetarisierung der Nutzerbeziehungen gegenüber Dritten stattfindet. Es könnte in dieser Hinsicht sinnvoll sein, beispielsweise die Gewinne, die ein digitales Geschäftsmodell aus der Monetarisierung der Nutzerdaten erwirtschaftet, in den Marktstaaten zu besteuern.

Die Intention, dass Steuern als Gegenleistung und Vergütung für die genutzte Infrastruktur in einem Staat entrichtet werden, wird jedoch mit der Besteuerung anhand des Unified Approach teilweise untergraben. Insbesondere bei der B-Ltd. wird deutlich, dass sowohl in den Niederlanden nur wenig bzw. in den Marktstaaten keine Infrastruktur genutzt wird, jedoch nach dem Unified Approach sehr wohl Besteueurungsrechte in diese Staaten verteilt werden. Dies bedeutet folglich auch, dass die Steuerbelastung auf den Bermudas reduziert werden würde, obwohl dort die Programmiertätigkeiten ausgeführt werden und somit auch Infrastruktur seitens der Mitarbeiter genutzt wird. Für Becker und Van der Ham [286] erweitert der Unified Approach die bisherige Einstellung, dass Steuern nicht nur für die Bereitstellung von Infrastruktur, sondern in der Zukunft auch für die Bereitstellung von Nutzerdaten und -informationen entrichtet werden.

Ein weiteres Problem bei der Besteuerung der digitalen Wirtschaft ist die Qualifizierung der Einkünfte. In Beispiel 1 muss die NL-B.V. an die B-Ltd. Lizenzzahlungen leisten. Die B-Ltd. ist auf den Bermudas unbeschränkt steuerpflichtig und durch die Lizenzeinkünfte auch in den Niederlanden beschränkt steuerpflichtig. Da die Niederlande laut nationalem Recht jedoch keine Quellensteuer auf Lizenzzahlungen erhebt, wird somit in diesem Beispiel keine Quellensteuer erhoben. Hingegen unterliegen Lizenzzahlungen in einigen Ländern, wie beispielsweise in Irland, aufgrund einer beschränkten Steuerpflicht dem Quellensteuer-Einbehalt. Dies würde umgelegt auf Beispiel 1 bedeuten, dass der gesamte Gewinn der B-Ltd. in den Niederlanden einer Quellensteuer unterliegen würde.

Insbesondere stellt sich bei dem Geschäftsmodell des Cloud-Computing die Frage, inwieweit das Übertragen der Rechte an einem Cloud-Computing-Geschäftsmodell als das Mieten eines Teils eines Servers (Lizenz) oder als Dienstleistung der B-Ltd. an die NL-B.V. (Unternehmensgewinn) quantifiziert werden kann. Diese Unterscheidung ist von grosser Relevanz, vor allem für die Anwendung von DBA, da das Besteuerungsrecht somit entweder dem Quellenstaat oder dem Ansässigkeitsstaat zugewiesen wird. In Bezug auf Fallbeispiel 1 bedeutet dies, dass die Zahlungen der NL-B.V. an die B-Ltd.

[283] Ebd., 18.

[284] Vgl. Bauer et al., „Die Besteuerung digitaler Wertschöpfung - "Doing nothing is not an option"," 889-890.

[285] Ditz und Pinkernell, „Neudefinition internationaler Besteuerungsrechte durch das OECD Inclusive Framework on BEPS - Eine Würdigung aus deutscher Sicht," 381.

[286] Becker und Van der Ham, „Das neue (digitale) ABC der Verlagerung von Besteuerungsrechten in die Marktstaaten," 2545.

nicht als das Mieten eines Teiles eines Servers, sondern als Dienstleistung der B-Ltd. an die NL-B.V. quantifiziert werden und diese Einnahmen nicht als Lizenzeinnahmen sondern als Unternehmensgewinne gelten würden. Folglich würden die Bermudas, als Ansässigkeitsstaat, das Besteuerungsrecht für die ganzen Einkünfte erhalten. Die Unternehmensgewinne würden somit nur auf den Bermudas zu einem Steuersatz von 0 % besteuert werden. Da in diesem Beispiel sowohl die Bermudas einen Steuersatz von 0 % aufweisen als auch die Niederlande keine Quellensteuer auf Lizenzzahlungen hat, ist die Unterscheidung irrelevant, da die Einkünfte nicht besteuert werden. Jedoch würde es in einem Land mit einer Quellensteuer, beispielsweise in Irland, einen wesentlichen Unterschied machen, wie Einkünfte qualifiziert werden.

4.2. Beispiel 2: Das traditionelle Geschäftsmodell der Unternehmensgruppe Y

Die Unternehmensgruppe Y ist, bis auf ein verändertes Geschäftsmodell, vollkommen identisch zur Unternehmensgruppe X. Das Geschäftsmodell der Unternehmensgruppe Y beinhaltet den Verkauf von Automobilen, wobei den Grossteil der Kunden Privatpersonen darstellen. Die verkauften Automobile liefern zudem Fahrdaten der Automobile an die Unternehmensgruppe Y, die diese auswertet und auch in ihrer weiteren Forschungs- und Entwicklungsarbeit nutzt. Die Unternehmensgruppe Y führt keine weiteren Geschäftstätigkeiten aus. Weiters übt sie eine wesentliche und auch länderspezifische Werbung in allen Marktstaaten aus.

Es wird in diesem Beispiel 2 nur auf die Thematik rund um den neuen Nexus, also das Fallen der Unternehmensgruppe Y in den Anwendungsbereich des Betrages A, näher eingegangen. Der Nexus der Beträge B und C knüpft immer noch am Konzept der physischen Präsenz an und wird hierbei nicht näher beleuchtet. Dieses Beispiel soll verdeutlichen, dass auch Geschäftsmodelle, welche nicht als digital bezeichnet werden können und Geschäftsmodelle der „Old Economy“ darstellen, vom Unified Approach und dessen Besteuerung betroffen sein können.

4.2.1. Mögliche Begründung eines Nexus für Betrag A

Im folgenden Abschnitt werden nun die unterschiedlichen Schwellenwerte (siehe 3.2.1.1), die ein Unternehmen überschreiten muss, um in den Anwendungsbereich des Unified Approach unter Betrag A zu fallen, anhand der Unternehmensgruppe Y näher betrachtet. Es wird überprüft, ob ein Teil des Gewinnes der Unternehmensgruppe Y in die Marktstaaten allokiert werden kann. Die Abbildungen 15 und 16 verdeutlichen diese Schwellen grafisch.

Die erste Schwelle, die ein Unternehmen überschreiten muss, ist ein konsolidierter Konzernumsatz von 750 Mio. EUR. Dies ist im Hinblick auf das Fallbeispiel 2 erfüllt, da die Unternehmensgruppe Y einen konsolidierten Konzernumsatz von 1.000 Mio. EUR aufweist.

Weiters muss die Unternehmensgruppe einen Aktivitäten-Test bestehen, um von Betrag A des Unified Approach betroffen zu sein. Hierbei geht es darum, ob die Unternehmensgruppe Y in die Gruppe der Automated digital services oder der Consumer-facing businesses eingeordnet werden kann. Dies kann anhand einer Analyse der Aktivitäten der Unternehmensgruppe Y gemacht werden. Eine Voraussetzung, um ein Consumer-facing business darzustellen, liegt in der Umsatzgenerierung aus dem Verkauf von Produkten bzw. Dienstleistungen. Weiters muss hierbei auch die Voraussetzung erfüllt sein, dass der Verkauf grundsätzlich an Konsumenten, also für private Zwecke, erfolgt. Der Konzern Y fällt aufgrund des Verkaufes von Automobilen an Privatpersonen in die Gruppe der Consumer-facing businesses.

Der Konzern Y fällt jedoch nicht nur durch den Verkauf von Automobilen in den Anwendungsbereich. Es müssen laut OECD noch weitere „Plus“-Faktoren definiert werden, die auf eine relevante und nachhaltige Interaktion des Unternehmens im Marktstaat hinweisen. In diesem Beispiel ist hierbei der Hinweis, dass die Unternehmensgruppe Y eine gezielte Werbung für die Kunden in den Marktstaaten schaltet, ein sogenannter „Plus“-Faktor. Hingegen ist der Hinweis, dass die Unternehmensgruppe Y Fahrdaten der Automobile auswertet und benutzt, nicht relevant für das Hineinfallen in den Anwendungsbereich. Auch ohne dieses Datenverwertung würde das Unternehmen in den Anwendungsbereich des Betrages A fallen. Dieser Hinweis der digitalen Komponente der Unternehmensgruppe Y ist somit irrelevant für die Begründung eines Nexus unter Betrag A, da bei den Consumer-facing businesses die „Plus“-Faktoren ausschlaggebend sind und nicht das Aufweisen einer digitalen Komponente.[287] Die Automobil-Branche wurde auch seitens der OECD explizit als ein Beispiel für ein Consumer-facing business genannt.[288]

Im nächsten Schritt müssen die Umsätze, die unter diese Consumer-facing activities fallen, eine gewisse Grösse aufweisen. In diesem Beispiel übt die Unternehmensgruppe Y keine weiteren Aktivitäten aus, weshalb der gesamte Unternehmensgewinn in die Gruppe der Consumer-facing activities fällt. Hierbei kann also davon ausgegangen werden, dass dieser de minimis-Schwellenwert vom Konzern Y übertreten wird.

Im Anschluss werden nun lediglich die Umsätze der Consumer-facing activities zur Begründung eines Nexus unter Betrag A herangezogen, deren Geschäftsbereich einen gewissen Prozentsatz an Rentabilität übersteigt. Die Höhe dieses Prozentsatzes wurde seitens der OECD noch nicht festgelegt. Es wird davon ausgegangen, dass die Unternehmensgruppe Y diese Rentabilität übersteigt.

Zudem muss die Summe der aggregierten Residualgewinne eine gewisse Höhe überschreiten, um eine Besteuerung anhand des Betrages A zu rechtfertigen. Auch dieser Betrag wurde von der OECD noch nicht festgelegt.

Zu guter Letzt folgt noch ein Nexus-Test für alle Marktjurisdiktionen, um die Allokation von Besteuerungsrechten in die Marktstaaten zu begründen. Diese Nexus-Bedingungen sind noch nicht präzise genug festgelegt. Die OECD spricht

[287] Vgl. OECD, „Outline of the Architecture of a Unified Approach on Pillar One,“ 12-13.
[288] Ebd., 11.

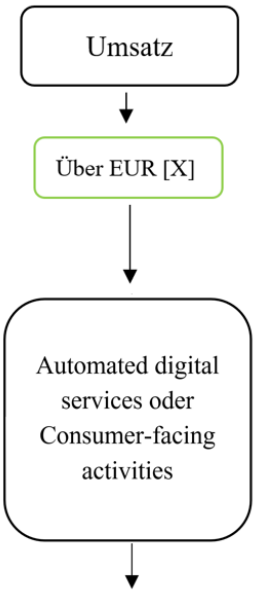


Abbildung 15: Nexus Betrag A – Teil 1
Eigene Abbildung

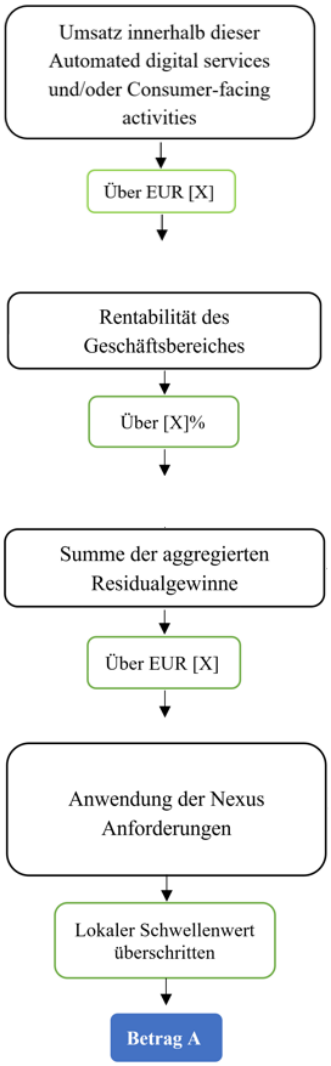


Abbildung 16: Nexus Betrag A – Teil 2
Eigene Abbildung

hierbei aber von einer signifikanten und nachhaltigen Rolle, die ein Unternehmen in einem Marktstaat ausüben muss. Als Beispiel wird die Generierung von Umsätzen in diesen Marktstaaten über Jahre hinweg genannt. Es wird die Annahme getroffen, dass die Unternehmensgruppe Y in allen Marktstaaten eine erhebliche Rolle ausübt und somit in den Niederlanden, Deutschland, Österreich und den USA mit dem Betrag A steuerpflichtig wird. In weiterer Folge wird nun der Betrag A berechnet und gemäss eines Umsatzschlüssels an die Marktstaaten verteilt, wie in Beispiel 1 dargestellt wird.

4.2.2. Bewertung des Unified Approach anhand der Unternehmensgruppe Y

Das zweite Beispiel der Unternehmensgruppe Y zeigt deutlich, dass der Unified Approach ein weit grösseres Ausmass annimmt, als nur die digitale Welt zu besteuern. Auch die „Old Economy“ kann durch diesen Unified Approach betroffen sein. Die sogenannten Consumer-facing businesses müssen keine digitalisierte Komponente mehr aufweisen, sondern eine signifkante und relevante Rolle in einem Marktstaat haben sowie weitere Plus-Faktoren, wie beispielsweise spezifische Werbung im Marktland, aufweisen. Dies reicht bereits aus, um in den Anwendungsbereich des Unified Approach zu fallen, ohne jegliche digitale Komponenten vorzuweisen. Auch Schön[289] weist darauf hin, dass der Unified Approach eine weit grössere Dimension erreicht, als nur die digitale Welt zu besteuern. Der Unified Approach in seiner neuesten Fassung zielt also nicht nur direkt auf die digitale Wirtschaft ab, sondern vor allem auf Geschäftsmodelle, die „remote“ funktionieren und auf sehr profitable Unternehmen.

Durch den breiten Anwendungsbereich können sehr unterschiedliche MNE mit dem Betrag A des Unified Approach

[289] Schön, „Seminar F (IFA/OECD): Internationale Besteuerung der digitalisierten Wirtschaft,“ 650.

besteuert werden. Die Heterogenität der Unternehmen sollte jedoch bei der Ausarbeitung des Unified Approach berücksichtigt werden. Wie bereits von Zinowsky[290] erwähnt, kann die Rentabilität der Geschäftsbereiche der Unternehmen höchst unterschiedlich sein. Insbesondere bei einem Vergleich von digitalen und traditionellen Geschäftsmodellen wie in Beispiel 1 und 2, können erhebliche Unterschiede der Rentabilität auftreten. Dies kann aber auch von Region zu Region sehr unterschiedlich sein, wie PwC[291] in ihrem Input-Statement zum Public Consultation Document beschreibt. Falls die Rentabilitätsschwelle seitens der OECD niedrig angesetzt wird, wie beispielweise zwischen 5 % und 7 %, werden ca. die Hälfte aller Unternehmen in allen Regionen der Welt in den Anwendungsbereich des Betrages A hineinfallen, so PwC. Folglich würde ein grosser Anteil des Besteuerungssubstrats in die Marktstaaten verschoben werden. Es stellt sich hierbei also auch die Frage, inwieweit diese Rentabtilitätsschwelle branchenabhängig bzw. regionenabhängig ausgearbeitet werden kann, sodass die Verschiedenartigkeit der Branchen und Regionen ausreichend beachtet wird.

4.3. Abschliessende Bewertung des Unified Approach

Die Analyse des Unified Approach mithilfe der Fallbearbeitung zeigt unterschiedliche Ergebnisse. In dieser abschliessenden Bewertung muss aber erwähnt werden, dass ein grosser Teil der Ausgestaltung des Unified Approach offen ist. Während einerseits viele Annahmen getroffen wurden, können andererseits noch wesentliche Änderungen am Unified Approach stattfinden. Die Analyse zeigt somit eine Momentaufnahme.

Der Unified Approach versucht die digitale Wirtschaft und ihre Geschäftsmodelle fair und effektiv zu besteuern. Dieser Ansatz der OECD ist sehr wohl in der Lage, BEPS-Probleme zu lösen, wenn gleich nicht eindeutig ist, wie viele neue BEPS-Probleme geschaffen werden. Auch das Konzept der physischen Präsenz kann durch den Unified Approach modernisiert und überwunden werden. Jedoch muss darauf aufmerksam gemacht werden, dass die Pauschalierung von digitalen Geschäftsmodellen ein Hauptproblem bei der Besteuerung derselben darstellt. Insbesondere den Wertschöpfungsbeitrag für alle Arten der digitalen Geschäftsmodelle zu identifizieren, ist nicht möglich. Jedoch kann ohne diese Analyse kaum geklärt werden, inwieweit Wertschöpfung in digitalen Geschäftsmodellen definiert werden kann und wie die User zur Wertschöpfung beitragen. Die Qualifizierung von Einkünften benötigt im Zeitalter der Digitalisierung neue Definitionen und kann durch den Unified Approach nicht direkt gelöst werden. Während es bei manchen Geschäftsmodellen im Anwendungsbereich sehr wohl sinnvoll erscheint, Besteuerungssubstrat in die Marktstaaten zu schieben, scheint es bei anderen Geschäftsmodellen unangebracht und nicht im Einklang mit der Wertschöpfung. Die Geschwindigkeit der Digitalisierung und Globalisierung nimmt immer mehr Fahrt auf, weshalb die digitalen Geschäftsmodelle in immenser Geschwindigkeit verändert werden und neue digitale Geschäftsmodelle entstehen.

Die Tabelle 12 fasst auf Grundlage der Fallbearbeitung zusammen, wie der Unified Approach zur Problembewältigung bei der Besteuerung der Digitalwirtschaft bewertet werden kann.

✓ Probleme konnten durch Unified Approach weitgehend gelöst werden

∼ Probleme konnten durch Unified Approach teilweise gelöst werden

× Probleme konnten durch Unified Approach nicht gelöst werden

Der Unified Approach wirft neben der Lösung für bestehende Probleme auch neue Fragen und Kritikpunkte auf, wobei die Hauptpunkte aus der Literaturanalyse nun nochmals kurz aufgegriffen und wenn möglich mit der Fallbearbeitung in Verbindung gebracht werden.

Im Hinblick auf die Fallbetrachtung kann gesagt werden, dass die Berechnung des Betrages A durch die standardisierte Formel bzw. die Berechnung des Betrages B durch die standardisierte Routinevergütung vereinfacht wird. Diese Meinung vertreten auch Pistone et al.[292] und Daurer[293], die damit eine Vereinfachung durch den Unified Approach sehen. Folglich können aufwendige Benchmark-Studien unterlassen werden. Jedoch kann bezüglich des Fallbeispieles 2 gesagt werden, dass sehr unterschiedliche Unternehmen, digitale Unternehmen wie auch Unternehmen der „Old Economy“, in den Anwendungsbereich des Unified Approach fallen. Diese Unternehmen können sehr unterschiedlich sein, weshalb eine standardisierte Routinevergütung unter Betrag B auch Nachteile nach sich ziehen kann. Es kann somit nicht auf individuelle Fälle und Besonderheiten eingegangen werden, so Becker und Van der Ham.[294] Eine zu hohe Routinevergütung für die Aktivitäten unter Betrag B in Betracht zum Median der Rentabilität würde ausserdem zu einer grossen Verschiebung des Besteuerungssubstrats in die Marktstaaten führen. PwC untersuchte hierbei die Schwankungen der EBIT-Margen von verschiedenen Regionen und Industrien, die teilweise grosse Unterschiede hervorbrachten. Für PwC scheint es somit schwierig, einen „one size fits all“-Prozentsatz unter Betrag B zu definieren.[295]

Weiters ist ein grosses Problem die Interaktion dieser drei Beträge und die daraus entstehende Komplexität, worauf Beer, Bulard und Najafi[296] aufmerksam machen. Dass zuerst

[290] Zinowsky, „Abkehr von tradierten Verrechnungspreismethoden und Einführung einer weltweiten Mindestbesteuerung,“ 814.

[291] PricewaterhouseCoopers, „Reaction to the OECD secretariat's consultation paper on the unified approach under Pillar 1,“ 18.

[292] Pistone et al., „The OECD Public Consultation Document “Secretariat Proposal for a 'Unified Approach' under Pillar One”: An Assessment,“ 15.

[293] Daurer, „Der Unified Approach der OECD,“ 556.

[294] Becker und Van der Ham, „Das neue (digitale) ABC der Verlagerung von Besteuerungsrechten in die Marktstaaten,“ 2544.

[295] PricewaterhouseCoopers, „Reaction to the OECD secretariat's consultation paper on the unified approach under Pillar 1,“ 27-30.

[296] Berr, Bulard und Najafi, „OECD-Vorschlag für eine einheitliche Besteuerung der digitalisierten Wirtschaft,“ 153.

Tabelle 12: Abschlussbewertung des Unified Approach

Probleme bei der Besteuerung von digitalen Geschäftsmodellen	Bewertung
Vereinfachung von Gewinnverschiebungs- und Gewinnverlagerungsmassnahmen	✓
Konzept der physischen Präsenz	✓
Komplexer Wertschöpfungsbeitrag	∼
Qualifizierung der Einkünfte und Zahlungen	×

die Beträge B und C anhand der Verrechnungspreisrichtlinien zugeteilt werden und im Anschluss Betrag A berechnet und verteilt wird, führt dazu, dass Betrag A eine Überlagerung darstellt. Dies sieht Daurer[297] als besonders kritisch, da dies zu einem Anstieg der Doppelbesteuerung führen würde. Dieses Problem kann auch in der Fallbearbeitung festgestellt werden. Die Gewinne und die Besteuerungsrechte dieser Gewinne unter Betrag A werden als Überlagerung auf die bisherigen Verrechnungspreise an die Marktstaaten verteilt. Bei einer Ansässigkeit der Muttergesellschaft in einem Hochsteuerland können jedoch Probleme der Doppelbesteuerung entstehen. PwC[298] weist im Hinblick auf diese Doppelbesteuerung unter Betrag A beispielsweise darauf hin, dass es möglich sein sollte, dass eine Gesellschaft ihre Bemessungsgrundlage um den Betrag, der den Marktstaaten zugeteilt wird, senken kann. Dies würde in diesem Beispiel dazu führen, dass die auf den Bermudas ansässige B-Ltd. ihre Bemessungsgrundlage um den Betrag senken kann, den die Marktstaaten erhalten haben. Eine andere Lösung, die von der OECD[299] in diesem Zusammenhang vorgeschlagen wird, ist die Anrechnung der Steuer, die in den Marktstaaten gezahlt wird. Hierbei stellt sich aber die Frage, wer steuerpflichtig für Betrag A ist und somit die Senkung der Bemessungsgrundlage bzw. die Anrechnung der gezahlten Steuer vornehmen kann. Zudem werden Konflikte bezüglich der Vereinbarkeit von Verrechnungspreisen und des Unified Approach als weiterer Problempunkt von Zinowsky[300] gesehen.

Ein weiteres Kernproblem, welches sich nicht durch die Fallbeispiele illustrieren lassen kann, ist die Notwendigkeit einer harmonisierten Gewinnermittlung.[301] Die Unterschiede zwischen nationalen und internationalen Rechnungslegungsstandards können gross sein und sind laut Schuster[302] nicht als Grundlage für eine gemeinsame Gewinnermittlung geeignet. Zudem wird die Ermittlung einer Bemessungsgrundlage auch durch die geplante Segmentierung von Geschäftsfeldern deutlich erschwert, so Daurer.[303] Diese ist nur mit einem sehr hohen Aufwand durchführbar. Hierbei erwähnen Greil und Wargowske[304] auch, dass die geplante Segmentierung in Geschäftsfelder zu neuartigen BEPS-Problemen führen kann.

Des Weiteren wurde bereits erwähnt, dass BEPS-Probleme durch den Unified Approach gelöst werden können. Dies ist auch insbesondere in Beispiel 1 ersichtlich, in welchem ein Konzern mit einer schädlichen Steuerstruktur durch den Unified Approach eine höhere Konzernsteuerquote aufweist. Die OECD[305] weist in der vorläufigen Analyse und Folgenabschätzung zu Säule 1 darauf hin, dass der Unified Approach teilweise Besteuerungsrechte von Niedrigsteuerländern in Länder mit höherer Steuerbelastung verschieben wird. Dies führt folglich zu leicht höheren globalen Steuereinnahmen. Laut Schön[306] lassen sich Märkte und User nicht einfach in Niedrigsteuerländer verschieben, weshalb BEPS-Probleme deutlich verringert werden können. Jedoch sind Greil und Wargowske[307] sowie auch Fuest[308] der Meinung, dass durch den sehr hohen Komplexitätsgrad des Unified Approach neue BEPS-Probleme geschaffen werden.

Wie Fallbeispiel 1 zeigt, wird die Steuerbelastung immer auf Grundlage des ganzen Konzernes berechnet. Wie bereits erwähnt, ist nicht klar, welche Gesellschaft schlussendlich steuerpflichtig für den Betrag A ist. PwC[309] bringt hierbei den Vorschlag ein, dass die Muttergesellschaft des MNE für den Betrag A steuerpflichtig wird, diesen berechnet und an die Marktjurisdiktionen zuweist. Weiters ist noch nicht geklärt, wie die Beträge steuerlich eingehoben werden. Dies sieht Daurer[310] vor allem in den Staaten als kritisch, in denen ein Konzern keine physische Präsenz besitzt, aber trotzdem mit Betrag A steuerpflichtig ist. Des Weiteren zeigt das Beispiel 1 auch, dass der Unified Approach nicht mehr nur eine Vergütung für die genutzte Infrastruktur darstellt. Durch das Aufkommen neuer digitaler Geschäftsmodelle wird diese Ansicht erweitert, sodass im Zeitalter der Digitalisierung Steuern auch für die Bereitstellung von Nutzerdaten und

[297] Daurer, „Der Unified Approach der OECD," 556.
[298] PricewaterhouseCoopers, „Reaction to the OECD secretariat's consultation paper on the unified approach under Pillar 1," 22-23.
[299] OECD, „Outline of the Architecture of a Unified Approach on Pillar One," 14.
[300] Zinowsky, „Abkehr von tradierten Verrechnungspreismethoden und Einführung einer weltweiten Mindestbesteuerung," 819.
[301] Ahlers, „Gründlichkeit vor Schnelligkeit," M4.
[302] Schuster, „Der "Unified Approach" der OECD - ein Überblick und erste Einschätzungen," 520.
[303] Daurer, „Der Unified Approach der OECD," 556.
[304] Greil und Wargowske, „Pillar 1 of the Inclusive Framework's Work Programme: The Effect on the Taxation of the Digi-tal Economy and Reallocation of Taxing Rights," 509.
[305] OECD, „Tax Challenges arising from the Digitalisation of the Economy," 13.
[306] Schön, „Seminar F (IFA/OECD): Internationale Besteuerung der digitalisierten Wirtschaft," 650.
[307] Greil und Wargowske, „Pillar 1 of the Inclusive Framework's Work Programme: The Effect on the Taxation of the Digi-tal Economy and Reallocation of Taxing Rights," 513.
[308] Fuest, „Die G20-Beschlüsse: Durchbruch zu einer besseren internationalen Besteuerung?," M5.
[309] PricewaterhouseCoopers, „Reaction to the OECD secretariat's consultation paper on the unified approach under Pillar 1," 4.
[310] Daurer, „Der Unified Approach der OECD," 556.

Nutzerinformationen gezahlt werden, so Becker und Van der Ham.[311]

Ein anderes relevantes Problem, welches für die Durchsetzung von wesentlicher Bedeutung sein wird, ist das Umgehen mit Verlusten. Hierbei wird seitens der OECD noch kein Lösungsvorschlag unterbreitet, weshalb dieses Problem auch nicht anhand der Fallbearbeitung demonstriert werden kann. Für Zinowsky[312] ist dieses Problem von grosser Bedeutung, da Staaten nicht damit einverstanden sind, im Gewinnfall Steuersubstrat abzugeben, während im Verlustfall der ganze Verlust alleine getragen werden muss.

Der Unified Approach soll den Ort der Besteuerung mit dem Ort der Wertschöpfung in Einklang bringen. Beispiel 1 zeigt, dass dieses Ziel durch den Unified Approach bei einigen Geschäftsmodellen nicht erreicht werden kann. Laut Müller und Bauerfeld[313] steht der Unified Approach nicht im Einklang mit einer Besteuerung am Ort der Wertschöpfung. Auch Bauer et al.[314] vertreten diese Ansicht. Ein Geschäftsmodell bereitet laut den Autoren nur steuerrechtliche Probleme, wenn der Ort der Besteuerung und der Wertschöpfung auseinanderfällt. Bei einem nicht Auseinanderfallen sind die bestehenden Besteuerungsregeln vollkommen ausreichend. Ein Auseinanderfallen dieser Orte tritt laut Bauer et al. nur dann auf, wenn drei Bedingungen kumulativ erfüllt sind: Dies ist eine Tätigkeit der Datenerhebung (1), wobei die erhobenen Daten einen wesentlichen Beitrag zur Monetarisierung leisten (2). Weiters darf keine Begründung eines Nexus im Marktstaat vorliegen (3). Hinsichtlich Fallbeispiel 1 würde das momentane Recht somit vollkommen ausreichen.

Ein weiteres Problem stellt neben den inhaltlichen Differenzen auch die Umsetzung des Unified Approach dar. Da der Unified Approach seine Wirkung laut Daurer[315] nur bei einer globalen Einführung entfalten kann, muss dies anhand eines globalen multilateralen Abkommens umgesetzt werden. Dies wird jedoch durch die Tatsache, dass viele Staaten des Inclusive Framework on BEPS keine DBA untereinander aufweisen, deutlich erschwert.

Der Ruf nach einer fairen Besteuerung für die Digitalwirtschaft hat zu einer geplanten Erneuerung geführt, die nicht nur die Digitalwirtschaft betrifft, sondern ein viel grösseres Ausmass annimmt, so Schön.[316] Die Stimmen sind zahlreich, dass der Unified Approach die Komplexität und auch die Doppelbesteuerung im internationalen Unternehmenssteuerrecht erhöhen wird. Dieser Kritik kann nach Ausarbeitung des Falles zugestimmt werden.

[311] Becker und Van der Ham, „Das neue (digitale) ABC der Verlagerung von Besteuerungsrechten in die Marktstaaten," 2545.

[312] Zinowsky, „Abkehr von tradierten Verrechnungspreismethoden und Einführung einer weltweiten Mindestbesteuerung," 813-814.

[313] Müller und Bauerfeld, „Risiken der Besteuerung der digitalen Wirtschaft nach „Säule 1"," 948.

[314] Bauer et al., „Die Besteuerung digitaler Wertschöpfung - "Doing nothing is not an option"," 889-890.

[315] Daurer, „Der Unified Approach der OECD," 556-557.

[316] Schön, „Seminar F (IFA/OECD): Internationale Besteuerung der digitalisierten Wirtschaft," 650.

5. Fazit

Digitalisierung und Globalisierung haben einen immensen Einfluss auf die Wirtschaft und das Geschäftsleben. Das grenzüberschreitende Handeln wird bedeutend vereinfacht und vielfach ist keine physische Präsenz in einem Absatzland mehr nötig, um Umsätze zu generieren. Dies führt jedoch auch dazu, dass gerade das Steuerrecht vor grosse Herausforderungen gestellt wird, da dieses vor allem auf dem Konzept der physischen Präsenz basiert. Es scheint Einigkeit dahingehend zu bestehen, dass das Konzept der physischen Präsenz im Zeitalter der Digitalisierung nicht mehr angemessen ist.

Weiters verschärft die Digitalisierung auch die Problematik rund um Gewinnverschiebungs- und Gewinnvermeidungsmassnahmen der multinationalen Unternehmen. Durch die steigende Abhängigkeit der Unternehmen von immateriellen Wirtschaftsgütern wird das Verlagern von Gewinnen in Niedrigsteuerländer deutlich vereinfacht, was sich als weiteres Problem der digitalen Welt abzeichnet. Gerade die grossen und bekannten Steuersparmodelle von Google & Co. führen immer wieder zu hoch politischen Diskussionen und daraus resultierenden Kämpfen rund um das Besteuerungssubstrat. Zudem entwickeln sich immer mehr digitale Geschäftsmodelle, bei denen der Wertschöpfungsbeitrag sehr schwer zu identifizieren ist. Die Wertschöpfungskette hat sich in der Digitalwirtschaft wesentlich verändert und es bleibt die Frage, inwieweit User zur Wertschöpfung eines digitalen Geschäftsmodelles beitragen. Es bedarf eines ausgeprägten Wissens über Geschäftsmodelle und ihrer Funktionsweise, um den Wertschöpfungsbeitrag zu ermitteln. Jedoch kann durch die Diversität, Komplexität und Vielzahl der unterschiedlichen digitalen Geschäftsmodelle dieser Wertschöpfungsbeitrag nur schwer für alle digitalen Geschäftsmodelle analysiert werden. Das BEPS-Projekt der OECD nennt das Ziel, eine Besteuerung am Ort der Wertschöpfung in Zukunft zu garantieren. Durch den immer grösser werdenden Druck der Politik arbeitet die OECD nun intensiv an einem neuen Ansatz, die Probleme, die durch die digitalisierte Welt entstehen oder verschärft werden, schnellstmöglich zu reduzieren oder gar zu lösen. Schon in den Anfängen des Projektes wurde seitens der OECD darauf aufmerksam gemacht, dass die digitale Welt nicht von der traditionellen Welt abgegrenzt werden kann. Deshalb beschränkt sich dieser Ansatz nicht nur auf rein digitale Unternehmen, sondern es können auch traditionelle Unternehmen, welche einen zunehmenden Grad an Digitalisierung aufweisen, davon betroffen sein.

Die Lösung der OECD, der Unified Approach, deutet auf eine umfassende und revolutionäre Änderung des internationalen Steuerrechts hin. Insbesondere die Nexus- und Gewinnverteilungsregeln erneuern das bestehende Steuerrecht grundlegend und verschieben somit Teile des Besteuerungssubstrats vom Ansässigkeitsstaat in den Nutzer- bzw. Marktstaat. Des Weiteren werden neue steuerliche Anknüpfungspunkte festgelegt, die jenseits des Erfordernisses einer physischen Präsenz existieren. Der Unified Approach sieht drei verschiedene Arten des Gewinnes vor, Betrag A, B und C, welche in die Marktstaaten verteilt werden können. Vor al-

lem bei Betrag A handelt es sich um eine steuerrechtliche Neuheit, da nun auch durch einen signifikanten und nachhaltigen Einfluss eines Unternehmens in einem Marktstaat ein Nexus begründet werden kann. Dieser Betrag A wird jedoch nur MNE betreffen, die einen Umsatz von 750 Mio. Euro aufweisen. Zudem gibt es noch weitere Schwellenbeträge, die erst überschritten werden müssen, um in den Anwendungsbereich des Betrages A zu fallen. Die Beträge B und C knüpfen an den bestehenden Verrechnungspreisregeln an und verfolgen das Ziel, diese zu vereinfachen. Während Betrag B eine feste Rendite für Basismarketing- und Vertriebsaktivitäten darstellt, vergütet Betrag C Aktivitäten, die über den Betrag B hinausgehen. Insbesondere das Zusammenspiel dieser drei Beträge muss weiter eruiert werden, um Doppelbesteuerung zu vermeiden und Rechtssicherheit zu gewährleisten.

Die Fallbetrachtung des Unified Approach zeigt unterschiedliche Ergebnisse. Während die BEPS-Probleme und das Problem der physischen Präsenz überwunden werden können, können die Probleme der komplexen Wertschöpfungsbeiträge und der Qualifizierung von Einkünften nicht eindeutig gelöst werden. Diese Probleme greifen weiter als der Unified Approach, da hier Definitionen im Zusammenhang mit der Digitalwirtschaft noch ausständig bzw. nicht geklärt sind. Der Unified Approach schafft es, einige Probleme bei der Besteuerung der digitalen Wirtschaft aus dem Weg zu schaffen und weist somit in die richtige Richtung. Dem Ruf nach einer höheren und gerechteren Besteuerung der Digitalwirtschaft wird mit dem Unified Approach Abhilfe geleistet. Es sind jedoch noch weitere Massnahmen nötig, um eine effektive Besteuerung zu gewährleisten und noch bestehende Probleme zu lösen.

Es bleibt offen, wie sich dieser Unified Approach in der Praxis umsetzen lässt, ohne eine immense Steigerung der Compliance- und Administrationskosten nach sich zu ziehen. Jedoch ist nach diesem neuen Entwurf des Unified Approach klar, dass es mit einer hohen Wahrscheinlichkeit eine grundlegende Änderung des internationalen Steuerrechts geben wird und die Auswirkungen aktuell noch unklar sind. Auch Unternehmen, die im Grunde nicht auf einem digitalen Geschäftsmodell basieren, können in den Anwendungsbereich fallen. Unternehmen, welche international tätig sind, ist dringend zu raten, eine Simulation der Konzernsteuerquote anhand der neuen Besteuerungsregeln durchzuführen, um so die Auswirkungen des Unified Approach zeitnah abschätzen zu können.

Der Grad der Digitalisierung wird auch in der Zukunft weiter zunehmen und die Digitalisierung wird weiterhin einen bedeutenden Einfluss auf den Erfolg eines Geschäftsmodelles haben. Durch den technologischen Fortschritt werden immer mehr digitale Geschäftsmodelle entwickelt. Falls der Vorschlag der OECD scheitern wird, werden mit Sicherheit weitere unilaterale Massnahmen zur Besteuerung der digitalen Wirtschaft eingeführt. Dies gilt es zu verhindern, um Doppelbesteuerungen zu vermeiden. Es bleibt mit Spannung abzuwarten, wie der nächste Entwurf des Unified Approach aussehen wird und ob ein globaler Konsens über ein derart politisch und wirtschaftlich relevantes Thema erreicht werden kann. Das internationale Steuerrecht ist im Wandel und die Veränderungen anhand des Unified Approach könnten grösser nicht sein. Auf diese Veränderung deuten auch Becker und Van der Ham[317] mit folgendem Zitat hin: „Die ABC-Regeln sind nicht einfach eine Evolution der bestehenden Besteuerungsprinzipien. Vielmehr handelt es sich um einen fundamentalen Paradigmenwechsel in der internationalen Unternehmensbesteuerung.“

[317] Becker und Van der Ham, „Das neue (digitale) ABC der Verlagerung von Besteuerungsrechten in die Marktstaaten,“ 2545.

Literatur

Ahlers, H. (2019). Gründlichkeit vor Schnelligkeit. *DB*, *72* (50), M4-M5.

Aigner, D., Bräumann, P., Kofler, G. & Tumpel, M. (2019). Die Veränderung des Steuerrechts durch die digitale Wirtschaft. In E. Felten, G. Kofler, M. Mayrhofer, S. Perner & M. Tumpel (Hrsg.), *Digitale Transformation im Wirtschafts- & Steuerrecht* (S. 305–320). Wien: Linde.

Allner, C. (o. J.). *300 Milliarden Euro: Steigender Anteil von Werbung in Sozialen Medien (inkl. Podcast)*. Zugriff am 07.03.2020 auf https://socialmediastatistik.de/werbung-social-media/

Arbeitskreis Steuern der Schmalenbach-Gesellschaft für Betriebswirtschaft e.V. (2018). Digitalisierung und Besteuerung. In S. Krause & B. Pellens (Hrsg.), *Betriebswirtschaftliche Implikationen der digitalen Transformation* (S. 123–142). Wiesbaden: Springer Fachmedien Wiesbaden.

Bauer, G., Fritz, J., Schanz, D. & Sixt, M. (2019). Die Besteuerung digitaler Wertschöpfung - "Doing nothing is not an option". *DStR* (16-17), 887–893.

Becker, J. & Englisch, J. (2017). Ein grösseres Stück vom Kuchen: Besteuerung der Gewinne von Google und Co. *Wirtschaftsdienst*, *97* (11), 801–808.

Becker, J. D. & Van der Ham, S. (2019). Das neue (digitale) ABC der Verlagerung von Besteuerungsrechten in die Marktstaaten: Neues OECD Konsultationspapier für die Weiterentwicklung der internationalen Unternehmensbesteuerung. *DB*, *72* (46), 2540–2545.

Bendlinger, S. (2017). Steht das Betriebsstätten-Konzept vor dem Aus? Möglichkeiten und Grenzen der Besteuerung der "Digital Economy". *WT Fachjournal* (3), 176–180.

Bendlinger, S. (2018). (Un-)Sinn der digitalen Betriebsstätte: Richtlinienpaket der EU zur Besteuerung der Digital Economy. *SWI*, *28* (6), 268–278.

Bendlinger, S. (2019). BEPS 2.0 - Besteuerung digitalisierter Geschäfte - eine (un)endliche Geschichte. *WT Fachjournal* (02), 108–110.

Berr, F., Bulard, C. & Najafi, N. (2020). OECD-Vorschlag für eine einheitliche Besteuerung der digitalisierten Wirtschaft. *Expert Focus* (3), 150–153.

Berr, F., Bulard, C. & Zahnd, N. (2019). Besteuerung der digitalisierten Wirtschaft. *Expert Focus* (9), 754–757.

Boehle, M. & Scholtholt, T. (2019). Gestaltungsoptionen zur gerechten Besteuerung von Digitalkonzernen: Eine kritische Diskussion zur aktuellen Arbeit des BEPS-Projekts der OECD. *IStR* (23), 919–924.

Daurer, V. (2019). Der Unified Approach der OECD. *ÖStZ* (20), 554–557.

Daurer, V. (2020). Unified Approach und Rechnungslegungsvorschriften - innovative Ideen für die Bemessung einer Ertragsteuer. *RWZ*, *10* (2), 51–55.

Deloitte. (2019). *Reaction to the OECD secretariat's consultation paper on the unified approach under Pillar 1.* London.

Ditz, X. & Pinkernell, R. (2019). Neudefinition internationaler Besteuerungsrechte durch das OECD Inclusive Framework on BEPS - Eine Würdigung aus deutscher Sicht. *ISR* (11), 377–389.

Duttiné, T. (2019). Lieber gut als schnell. Keine populistischen Lösungen für die Besteuerung der digitalen Wirtschaft. *BB* (48), 1.

Ehlermann, C. (2019). Welche Richtung nimmt die Besteuerung digitaler Geschäftsmodelle? *DB* (15), M4-M5.

Faulhaber, L. V. (2019). Taxing Tech: The Future of Digital Taxation. *Virginia Tax Review* (145), 1–52.

Felten, E., Kofler, G., Mayrhofer, M., Perner, S. & Tumpel, M. (Hrsg.). (2019). *Digitale Transformation im Wirtschafts- & Steuerrecht* (1. Aufl.). Wien: Linde.

Fuest, C. (2019). Die G20-Beschlüsse: Durchbruch zu einer besseren internationalen Besteuerung? *DB* (29), M4-M5.

G7. (2019). *Working for a fairer Capitalism: Press Kit: Chantilly 17-18 July 2019.*

Grassl, B. & Koch, R. (2019). Unilaterale Initiativen zur Einführung von Digitalsteuern. *IStR*, *28* (22), 873–888.

Greil, S. & Wargowske, L. (2019). Pillar 1 of the Inclusive Framework's Work Programme: The Effect on the Taxation of the Digital Economy and Reallocation of Taxing Rights. *BIT*, *73* (10), 504–513.

Heidecke, B., Lappe, S., Pandeya-Koch, A. & Wilmanns, J. (2019). Quo vadis Verrechnungspreise - die OECD auf Abwegen? Der Entwurf zur "Säule 1" v. 9.10.2019. *IWB* (23), 931–940.

Junkers, L. (2018). Lösungen der OECD/G20 und Europäischen Kommission zur Besteuerung der Digital Economy - Eine kritische Auseinandersetzung. *ISR* (10), 370–378.

Kempelmann, G. (2019). The Future of International Tax. *IStR* (16), 662–665.

Kirchmayr, Mayr, Hirschler, Kofler & Ehrke-Rabel (Hrsg.). (2018). *Digitalisierung im Konzernsteuerrecht: Schriftenreihe zum Konzern- und Unternehmenssteuerrecht*. Linde.

Kofler, G., Mayr, G. & Schlager, C. (2017). Digitalisierung und Betriebsstättenkonzept. *RdW* (5b), 369–388.

Krause, S. & Pellens, B. (Hrsg.). (2018). *Betriebswirtschaftliche Implikationen der digitalen Transformation*. Wiesbaden: Springer Fachmedien Wiesbaden.

Kreienbaum, M. (2019). Fortschritte bei der Digitalbesteuerung - Zweisäulenstrategie in der Diskussion. *IStR* (4), 121–123.

Ludwig, C., Sprengel, C. & Klein, D. (2019). Eine Digitalsteuer gefährdet Wettbewerbsfähigkeit und Rentabilität. *ZEW News* (12), 7.

Mahr, B. (2019). Digitalsteuer - kein Spielfeld für nationale Alleingänge. *GWR* (11), 189–191.

Mayr, G. & Schlager, C. (2018). Digitale Geschäftsmodelle und Betriebsstättenkonzept. In Kirchmayr, Mayr, Hirschler, Kofler & Ehrke-Rabel (Hrsg.), *Digitalisierung im Konzernsteuerrecht* (S. 21–32). Linde.

Müller, S. & Bauerfeld, T. (2019). Risiken der Besteuerung der digitalen Wirtschaft nach „Säule 1“: Handlungsbedarf des deutschen Gesetzgebers. *IWB* (23), 941–951.

OECD. (1999). *A Global Action Plan for Electronic Commerce: Prepared by Business with Recommendations and for Governments* (Bd. 44). Paris: OECD Publishing.

OECD. (2014). *Aktionsplan zur Bekämpfung der Gewinnverkürzung und Gewinnverlagerung*. Paris: OECD Publishing.

OECD. (2015a). *Herausforderungen für die Besteuerung der digitalen Wirtschaft*. Paris: OECD Publishing.

OECD. (2015b). *Measuring and Monitoring BEPS, Action 11 - 2015 Final Report: OECD/G20 Base Erosion and Profit Shifting Project*. Paris: OECD Publishing.

OECD. (2018a). *OECD-Verrechnungspreisleitlinien für multinationale Unternehmen und Steuerverwaltungen 2017*. Paris: OECD Publishing.

OECD. (2018b). *Steuerliche Herausforderungen der Digitalisierung - Zwischenbericht 2018: Inclusive Framework on BEPS: OECD/G20 Projekt Gewinnverkürzung und Gewinnverlagerung*. Paris: OECD Publishing.

OECD. (2019a). *Addressing the Tax Challenges of the Digitalisation of the Economy - Policy Note: 23 January 2019.*

OECD. (2019b). *Addressing the Tax Challenges of the Digitalisation of the Economy: Public Consultation Document: 13 February - 6 March 2019.*

OECD. (2019c). *OECD Secretary-General Tax Report to G20 Finance Ministers and Central Bank Governors: October 2019.* Paris.

OECD. (2019d). *Secretariat Proposal for a "Unified Approach" under Pillar One: Public Consultation Document: 9 October 2019 - 12 November 2019.*

OECD. (2020a). *Coronavirus (Covid-19): Update on OECD tax work.* Zugriff am 31.03.2020 auf http://www.oecd.org/tax/beps/coronavirus-covid-19-update-on-oecd-tax-work.htm

OECD. (2020b). *Outline of the Architecture of a Unified Approach on Pillar One: 29-30 January 2020.* Paris.

OECD. (2020c). *Tax Challenges arising from the Digitalisation of the Economy: Update on the Economic Analysis & Impact Assessment: Webcast 13 February 2020.*

Olbert, M. & Sprengel, C. (2019). Taxation in the Digital Economy: Recent Policy Developments and the Question of Value Creation: Working Paper. *ZEW Discussion Paper* (19-010), 1–31.

Pistone, P., Pinto Nogueira, J. F., Andrade Rodríguez, B. & Turina, A. (2020). The OECD Public Consultation Document "Secretariat Proposal for a 'Unified Approach' under Pillar One": An Assessment. *BIT*, *74* (1), 14–27.

Postler, S. (2019). Die Digitalisierung traditioneller Geschäftsmodelle: Ein Ansatz zum Umgang mit steuerlichen Verrechnungspreisen. *IStR* (3), 100–107.

PricewaterhouseCoopers. (2019). *Reaction to the OECD secretariat's consultation paper on the unified approach under Pillar 1.* London.

Pross, A. (2013). *Faire Besteuerung multinationaler Unternehmen: Neue Initiativen internationaler Steuerpolitik.* Berlin.

Scherff, D. (2014). *Millionen in Millisekunden: Hochfrequenzhandel.* Zugriff am 20.02.2020 auf https://www.faz.net/aktuell/finanzen/fonds-mehr/hochfrequenzhandel-millionen-in-millisekunden-12803798-p2.html

Schneemelcher, P. & Dittrich, P.-J. (2019). *Nutzer, Daten, Netzwerke: Ein Vorschlag zur Besteuerung der Digital-Wirtschaft im europäischen Binnenmarkt.* Berlin.

Schön, W. (2019). Seminar F (IFA/OECD): Internationale Besteuerung der digitalisierten Wirtschaft. *IStR* (16), 647–650.

Schuster, J. (2019). Herkulesaufgabe Digitalisierung: Herausforderungen an das EU-Recht: Gastkommentar. *EuZW*, *30* (24), 1001–1002.

Schuster, S. (2019). Der "Unified Approach" der OECD - ein Überblick und erste Einschätzungen. *SWI*, *29* (11), 514–521.

Stévenard, J.-P. (2019). Der Entwurf einer französischen Digitalsteuer. *IWB* (7), 270–272.

Szer, B. (2013). *Cloud Computing und Wissensmanagement: Bewertung von Wissensmanagementsystemen in der Cloud*. Hamburg: Diplomica Verlag.

Tørsløv, T., Wier, L. & Zucman, G. (2019). *40% of multinational profits are shifted to tax havens each year.* Zugriff am 03.04.2020 auf https://missingprofits.world/

Versin, V. & Hidien, J. W. (2019). Digitales Geschäftsmodell als steuerlicher Wettbewerbsfaktor. *GmbHR* (14), 759–770.

Vornholz, G. (2019). *Digitalisierung der Immobilienwirtschaft*. Berlin: De Gruyter.

Weggenmann, H., Blank, A. & Brunnhübner, A. (2019). OECD Public Consultation Document/Programme of Work betreffend Vorschläge zur Besteuerung der Digitalwirtschaft - schöne neue Steuerwelt? *IStR*, *28* (19), 769–776.

Wenz, M. (2020a). *The Future of International Taxation: Digitalization and Minimum Taxation - Behavioural Responses - Cases - Lobbying on BEPS Project: February 2020.*

Wenz, M. (2020b). *Internationale Unternehmensbesteuerung: Holding- und Finanzierungsgesellschaften.*

Wünnemann, M. (2019). Herausforderungen der Besteuerung der digitalen Wirtschaft im Jahr 2019. *IStR* (4), 134–140.

Zinowsky, T. (2019). Abkehr von tradierten Verrechnungspreismethoden und Einführung einer weltweiten Mindestbesteuerung: Diskussion des Arbeitsprogramms für die Verabschiedung einer neuen "Weltsteuerordnung". *IStR* (20), 811–820.