JUMS

Volume 6, Issue 3, September 2021

Advisory Editorial Board:

DOMINIK VAN AAKEN
FREDERIK AHLEMANN
BASTIAN AMBERG
THOMAS BAHLINGER
CHRISTOPH BODE
SULEIKA BORT
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
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
TILL TALAULICAR
ORESTIS TERZIDIS
ANJA TUSCHKE
MATTHIAS UHL
STEPHAN WAGNER
BARBARA E. WEIßENBERGER
ISABELL M. WELPE
HANNES WINNER
THOMAS WRONA
THOMAS ZWICK

JUNIOR MANAGEMENT SCIENCE

Max Lindner, Sustainable Food Consumption of German Millennials: Exploring the "Attitude Behavior Gap" 424

Nele Braun, The Role of the European Central Bank in a Sustainable Financial System 468

Elif Gürlek, Corporate Taxation in the U.S. and Canada – A Comparative Analysis 489

Markus Probst, Ohne Moos nichts los? Eine Discrete-Choice-Analyse der Präferenzen von Medizinstudenten für die Arbeit auf dem Land 507

Tim Leonard Hermes, The Impact of Community Involvement on Game Life-Cycle: Evidence based on Gaming Platform Steam 547

Leonie Rinke, Grenzüberschreitende ertragsteuerliche Organschaft 568

Konstantin Spring, Backtesting the Expected Shortfall 590

Victoria Roth, Talking is Silver, Doing is Gold? – The Influence of Corporate Social Responsibility on Corporate Financial Performance 637

Published by Junior Management Science e.V.

Junior Management Science 6(3) (2021) 424-467

Junior Management Science

journal homepage: www.jums.academy

Sustainable Food Consumption of German Millennials: Exploring the "Attitude Behavior Gap"

Max Lindner

Ludwig-Maximilians-Universität München

Abstract

This thesis picked up on the recent trend of sustainability and sustainable development by investigating the sustainable food consumption of German Millennials regarding an assumed gap between their attitude and their actual consumer behavior. Based on the Theory of Planned Behavior, a multi-step, qualitative research process was developed in order to find explanations for the attitude behavior gap. The process triangulated data from questioning ten German Millennial consumers via qualitative interviews, observing these consumers during a regular grocery shopping trip and analyzing the products they bought. The study revealed the following barriers preventing the participants from consuming according to their attitude: price, lack of knowledge on part of the consumers, insufficient sustainable product ranges, consumers minimizing the time spent for grocery shopping, habitualized purchase decisions and in exceptional cases a desire for unsustainable products. Accordingly, recommendations were derived for sustainable food manufacturers and supermarket chains, public policy and consumers. Additionally, this thesis contributes to consumer research on the attitude behavior gap as well as on behavioral research by applying and extending the Theory of Planned Behavior.

Keywords: Sustainable consumption; attitude behavior gap; consumer behavior; food consumption Verbal Protocol Analysis; Theory of Planned Behavior.

1. Introduction

1.1. Relevance and Scope

The Need for Sustainable Development and Sustainable Consumption

The world faces multiple environmental problems like drastic changes in weather patterns or an increased number of natural disasters, mostly caused by men due to the emission of greenhouse gases (GHG), especially carbon dioxide (C02). Pollution through manufacturing, deforesting and more environmentally harmful activities increased the C02 concentration by more than 33 percent (Trudel, 2019). Despite various counteractions, the mass of C02 emitted is far from declining. The worldwide annual emissions went up by 2.13% from 35.81 million tons in 2017 to 36.57 million tons in 2018 (Global Carbon Project, 2019). Besides environmental problems, other issues like preserving health or fighting poverty get more and more important (Bernyte, 2018).

In this context, sustainable development is gaining more and more attention. It was first popularized in 1987 by the World Commission on Environment and Development (WCED) when they published “Our common future”, also known as the “Brundtland Report” (Atapattu, 2019). In that report, the WCED acknowledged that both environmental protection and economic development were important, as they claim poverty is the biggest polluter and economic development is necessary for developing countries to improve the living standards of their people. Therefore, sustainable development was defined as a “development that meets the needs of the current generation while not comprising the needs of future generations” (Brundtland, Khalid, Agnelli, Al-Athel, & Chidzero, 1987, n.p.).

More recently the initiative of the Sustainable Development Goals was introduced in 2012 by the United Nations (UN). These goals are globally acknowledged and were officially adopted by the different states in 2015 (Hák, Janoušková, & Moldan, 2016; Sachs, 2012). The SDGs look beyond narrow economic measures of progress and consider all aspects of well-being for current and future generations, especially regarding the environment but also to eradicate

DOI: https://doi.org/10.5282/jums/v6i3pp424-467

poverty worldwide and to safeguard the planet (United Nations, 2015). Since then, sustainable development is defined in more detail by the 17 SDGs shown in figure 1. Each one of them is addressing another current challenge, however, they are integrated and indivisibly connected (Atapattu, 2019; Hák et al., 2016).

Integrating the three main and interlinked dimensions of economic, social, and environmental development, sustainable development has become the center of a renewed development framework for countries to meet the changing development priorities and gaps that previous strategies have been unable to close (Verma, Petersen, & Lansford, 2019).

According to Hanss, Böhm, Doran, and Homburg (2016), there is a common census that besides institutions and profit-driven firms, also single individuals have some responsibility to contribute to sustainable development by consuming sustainably. Łuczka and Smoluk-Sikorska (2017) even state that this sustainable consumption is one of the main elements of sustainable development. The United Nations, n.d.-b even formulated one specific SDG about consumption, which shows its high relevance.

The Apparent Gap Between Attitude and Behavior Regarding Sustainable Consumption

Consumer awareness regarding sustainability has grown over the past years and sustainability gets more and more important in the academic literature (Carrigan, Szmigin, & Wright, 2004; Connolly & Shaw, 2006; Dutta, 2014; Wei, Chiang, Kou, & Lee, 2017; Wiederhold & Martinez, 2018). Ellen, Wiener, and Cobb-Walgren (1991) point out that consumers' environmental concerns and the awareness that their own actions impact directly the environment drive sustainable attitude. Wei et al. (2017) also state that consumers are aware that their buying behavior and consumption patterns have an impact on environmental problems like pollution and climate change. According to Bellmann and Koch (2019), more than 50% of the questioned companies state ecological sustainability as important or even (very) important in their business activity. On the other hand, more than 60% of private consumers in Germany claimed to have changed their consumer behavior due to ecological concerns (Boyon, 2020). This indicates that many people hold a positive attitude towards sustainability and sustainable consumption.

Despite the consumers' awareness and interest in environmental as well as in social challenges and a resulting positive attitude, this does not always lead to sustainable behavior, captured by actual purchase decisions in this context. Prothero et al. (2011) claim that even though 40% of consumers are willing to buy environmentally friendly, only 4% actually do so. Recent literature has already picked up on these diverging observations, indicating the existence of this gap for different products and geographical locations (K. Lee, 2008; Moraes, Carrigan, & Szmigin, 2012; Wiederhold & Martinez, 2018; Young, Hwang, McDonald, & Oates, 2010).

General Differences in Consumer Behavior

Prior literature suggests that consumers of different nations and therefore different cultural backgrounds show considerable differences in consumer behavior (Gentina, Butori, Rose, & Bakir, 2014; Gómez-Luciano, de Aguiar, Vriesekoop, & Urbano, 2019; Kacen & Lee, 2002; Lapierre & Rozendaal, 2018; Manrai, 2018). For example, Gentina et al. (2014) state that susceptibility to peer influence drives teenage consumer shopping in the Netherlands, while in the USA it is additionally motivated by the need for uniqueness. Another example is the work of Gómez-Luciano et al. (2019), which suggests that countries from higher economically developed countries tended to show more readiness to replace traditional meats with alternatives.

In the context of the thesis, it is sensible to choose Germany as the focus country, as Germany has the fourth-largest Gross Domestic Product worldwide which shows its enormous economic relevance (World Population Review, 2019). Additionally, sustainability has high relevance for the German government (Bundesregierung, 2019) as well as for the public and population, which makes it a sensible country to look at in the context of sustainable development (Umweltbundesamt, 2017, 2019a). Furthermore, the researcher is based in Germany and can easily access German interviewees, probands, or respondents.

Similar to the national and cultural background, the generational affiliation seems to influence consumer behavior, as suggested by current literature (Bulut, Kökalan Çımrin, & Doğan, 2017; Miller & Washington, 2019; Priporas, Stylos, & Fotiadis, 2017; Southgate, 2017). Therefore, it is reasonable to focus on one generational group. Members of Generation Y, also called Millennials, are said to have a more environmentally friendly attitude than older generations (Bedard & Tolmie, 2018) as they are the first consumers to grow up in a globally interdependent world (Bucic, Harris, & Arli, 2012). Furthermore, their presence in the workforce and their spending power rises (The Nielson Company, 2015).

The Food Industry's Major Role in Sustainable Consumption

The German food industry covers the production and processing of all kinds of food and beverages[1] and is a major part of the economy (Bundesministerium für Wirtschaft und Energie, 2019). The mostly small and medium-sized 6,119 companies in the industry employ 608,553 people and generated a revenue of 179.6 billion Euro in 2018 (+0% growth from 2017). That means that every seventh industrial company in Germany is producing food, employing 9.5% of all employed persons in Germany. Furthermore, Germany is the third-largest exporter of food worldwide with an export ratio -of 33.5% (Bundesverband der deutschen Ernährungsindustrie, 2019).

On the other hand, the production and consumption of food are responsible for 20% to 30% of the environmental impacts in Germany (Umweltbundesamt, 2013). According to Grünberg, Nieberg, and Schmidt (2010), emissions for the food industry in Germany, including processing, transport and distribution, comprises about 16% to 22% of the total emissions of GHGs. Osterburg et al. (2009) state that other authors attribute between 14% and 22% of the emissions to

[1] See the appendix section A1.2 for a detailed overview of the German food industry.

(United Nations, n.d.-b)

Figure 1: Overview of the 17 Sustainable Development Goals

the food industry. Tukker et al. (2006) claim that food and drink cause 20% to 30% of the various environmental impacts of private consumption (in the EU). In a more recent investigation, the World-Resource-Institute (2016) estimates that the total food chain impact on GHGs emissions is 25% to 30% (worldwide).

The environmental impacts include not only GHGs but also other environmental pollutants. They acidify the soil or water or destroy the ozone layer. The production and processing of food also requires many resources. Water is of great importance to the food industry. Due to the processing steps and hygiene requirements, the water demand in the food industry is high. Furthermore, wastewater from food manufacturing is often heavily polluted with organic substances. Also, the drying, cooking, and cooling of food is energy-intensive. This energy consumption burdens resources as well as the climate. Energy efficiency is therefore an important lever when implementing the best available technology (Umweltbundesamt, 2013).

Also, from a social aspect, the food industry is interesting. In recent years many big food manufacturers and retailers were involved in various scandals ranging from water exploitation in Africa (Handelsblatt, 2013) over violations of animal rights, contaminated food (PETA, 2019), and child labor (Flatley, 2019).

Summary

Due to various environmental and social challenges, sustainable development gained relevance in the last years. As a major lever for individual consumers, sustainable consumption gets more and more important as well. Even though consumer attitudes towards sustainability increases, it is not reflected in actual consumer behavior, indicating a gap. Closing this gap is of high relevance in the context of sustainable development. The food industry plays a major role, economically as well as ecologically and socially, so looking closer at that industry is sensible. Due to considerable differences in consumer attitude and behavior dependent on the national and cultural background as well as on the affiliation of a generational group it is promising to focus on a specific country and a specific generation. Reasonable choices for the country and the generation are Germany and the Millennials.

1.2. Goal and Contribution

The goal of this thesis is to explore the gap between consumers' attitude and the behavior and to provide specific reasons respectively explanations for it. In doing so, this thesis contributes to the superior goal, which is closing the attitude behavior gap (ABG) and thereby fostering sustainable consumption. By providing insights regarding consumer behavior, implications for sustainable food manufacturers and supermarket chains (e.g. improving marketing and sales activities) and public policy (e.g. development of purposeful and effective regulations) can be derived as well as implications for the consumers themselves (e.g. revelation of unconscious beliefs).

Additionally, this thesis is also contributing to the current state of consumer research regarding the attitude behavior gap as well as to general behavioral research.

As implied by the previous section, this thesis is limited to the consumption of food and does not include the consumption of any other kind of product. Also, it is restricted to Germany, respectively German consumers and focuses only on the consumption behavior of Millennials. Therefore, citizens from other countries as well as consumers with a different generational affiliation are excluded.

So far, the consumer behavior of German Millennials regarding sustainable food consumption has not been investigated despite the market's enormous economic (Bundesmin-

isterium für Wirtschaft und Energie, 2019), environmental (Grünberg et al., 2010; Osterburg et al., 2009; Umweltbundesamt, 2013) and social relevance (Flatley, 2019; Handelsblatt, 2013; PETA, 2019).

Therefore, this thesis tries to examine the following research question:

> How can the assumed attitude behavior gap regarding the sustainable food consumption of German Millennials be explained, respectively what are specific reasons for the gap?

1.3. Structure

To guide the reader through this thesis, it has been divided into six main parts. These parts contain several sections, which again contain chapters. In some cases, chapters are additionally divided into sub-chapters.

The first part introduces the overall topic and shows its great relevance in public discussion and research. The second part defines the basic concepts of this thesis like sustainable consumption as well as the German Millennials and sustainable food. After that, the Theory of Planned Behavior (TPB) is introduced as the guiding theory of this thesis explaining consumer behavior, before the second part is concluded with a summary and the resulting research question. The third part develops a detailed qualitative study design, triangulating qualitative interviews, observations, and the analysis of actual purchases, in order to examine the research question. This is followed by the part that presents and interprets the results of the study and afterwards the thesis is critically discussed, and limitations are examined in the fifth part. At last, the thesis is concluded with part six, which describes the theoretical contributions and possible further research and practical implications that can be derived. The structure of the thesis is displayed below in figure 2.

2. Conceptual Framework

2.1. Basic Concepts, Definitions, and Distinctions

This first section of this part defines the main concepts of this thesis as a foundation for the further process.

2.1.1. Definition of Consumption

Consumption is a term that is commonly used in daily life and in academic literature, but it is used in a different context with different meanings (Stern, 1997). In the context of economics, it means spending funds for utility (Black, Hashimzade, & Myles, 2012; D'Orlando & Sanfilippo, 2010), while in biology it means eating and digesting other members of a food chain (Biology Dictionary, n.d.). The Cambridge Dictionary (2014) provides a rather practical and daily life definition for consumption: "the amount used or eaten", respectively "the act of using, eating, or drinking something". Another online dictionary confirms these definitions and adds one more particular definition: "the act of buying and using things" (Collins Cobuild, n.d.). From a general consumers' perspective, consumption is defined as a three-stage process that involves purchasing, using, and disposing goods and/or services (Halkier, 2001; Kim et al., 2012; Stern, 2000).

This thesis deals with a specific form of consumption from a consumer's perspective, the consumption of food. Therefore, it is obvious that using definitions from economics, biology or other unrelated fields is not sensible and a consumer perspective's definition should be used. As this thesis aims to elicit insights about purchasing behavior, it is sensible to use a definition that also includes the act of purchasing. Therefore, in accordance with Collins Cobuild (n.d.); Halkier (2001); Kim et al. (2012); Stern (2000), the following definition is used for this thesis:

> Consumption is the act of purchasing, using, and disposing goods.

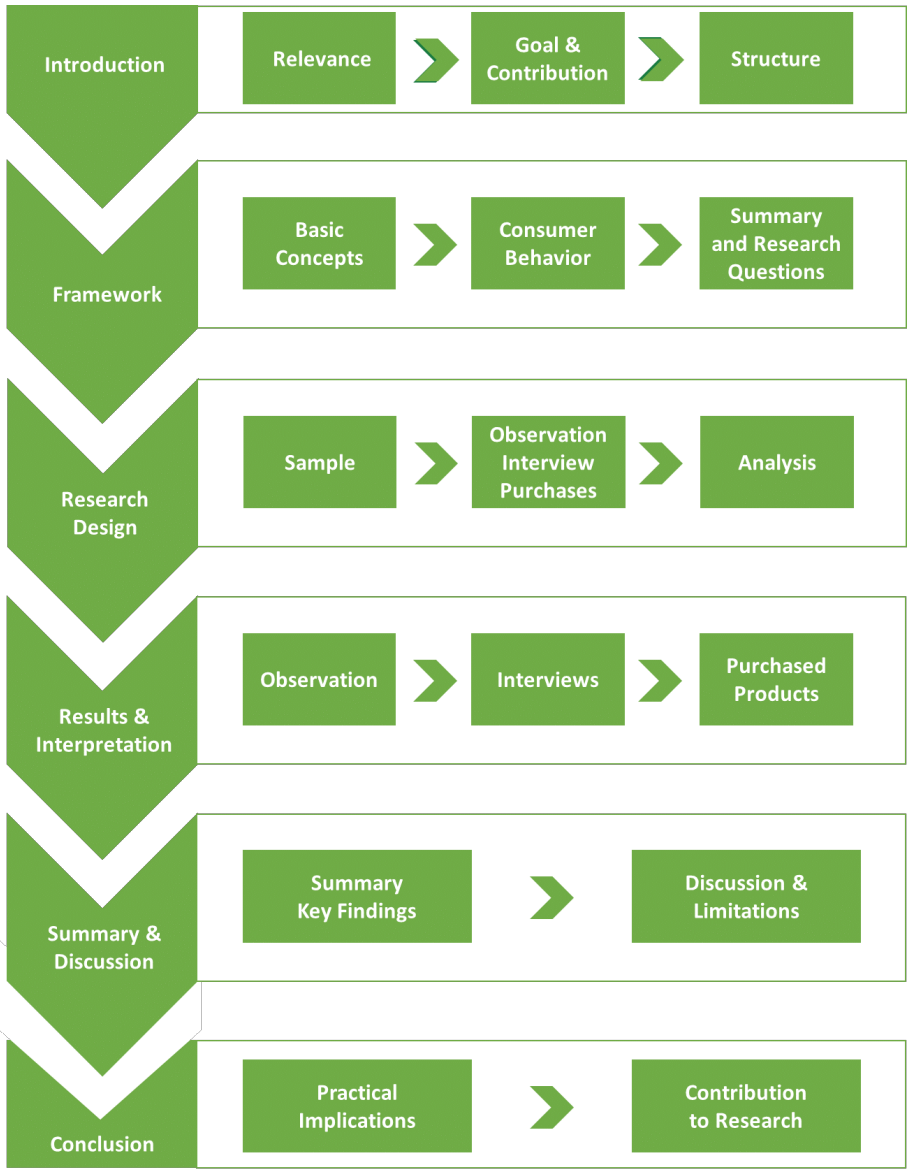
2.1.2. Definition of Sustainable Consumption

In literature, there is no census regarding the definition of sustainable consumption. Even though definitions are similar, authors include different aspects according to Łuczka and Smoluk-Sikorska (2017). Some declare sustainable consumption as pro-environmental consumption (Brough, Wilkie, Ma, Isaac, & Gal, 2016; Chen, Ghosh, Liu, & Zhao, 2019), meaning the consumption of products that are produced in a way that reduces the environmental impact as much as possible. Others, like Łuczka & Smoluk-Sikorska, 2017 or Scott & Weaver, 2018, do also include social aspects, e.g. exclusion of particular social groups, fair trade, animal rights and so an. Often the term is used without clearly defining it (Luchs, Phipps, & Hill, 2015; Prothero et al., 2011).

The UN's Definition of Sustainable Consumption

While there is no census about the definition of sustainable consumption in academic literature, a definition can also be derived from the UN. In 1994, the Oslo Symposium, an initiative of the United Nations Conference on Environment that focused on sustainable consumption and production (Baker, 1996), firstly defined sustainable consumption as the use of products and services which respond to basic needs and improve the quality of life while minimizing the use of natural resources and toxic materials as well as waste and pollutants over the whole product or service life cycle. (Sustainable Development Knowledge Platform, n.d.).

Later, the UN formulated, as previously mentioned, one specific SDG about sustainable production and consumption. By doing so, the UN points out the strong link between them. In fact, sustainable consumption is nothing else than the consumption of sustainable products respectively products that result from sustainable production. Sustainable production aims at decoupling economic growth from environmental degradation, increasing resource efficiency, and promoting sustainable lifestyles by reducing future economic, environmental, and social costs. Therefore, sustainable consumption and production should contribute substantially to poverty alleviation and the transition towards low-carbon and green



Own representation

Figure 2: Structure of the thesis

economies (United Nations, n.d.-a; United Nations Environment Programme, 2015).

Synthesized Definition for Sustainable Consumption

The SDGs represent common goals for the members of the UN and were officially adopted by the different states in 2015 (Hák et al., 2016; Sachs, 2012). Therefore, they are a suitable basis for a definition. For this thesis, the following definition will be used:

> Sustainable consumption is the consumption of products resulting from sustainable production. Sustainable production is a form of production that reduces environmental degradation, increases resource efficiency, and promotes sustainable lifestyles compared to conventional forms of production. Furthermore, it contributes to Sustainable Development by benefitting one or more related SDGs. In summary sustainable production reduces future economic, environmental and social costs.

Term Distinction

A review of the academic literature shows that besides sustainable consumption similar forms of consumption are mentioned. Namely, these terms are "green consumption", "responsible consumption" and "ethical consumption". There is no general census about what aspects these different types of consumption cover (social or environmental aspects or both) and the terms are not properly differentiated. Therefore, a literature review[2] was conducted in order to distin-

[2]See the appendix section A.1.3 for the research strategy and additional

guish the different forms of consumption. As it is not the goal of this thesis to do a fully comprehensive literature review about sustainable consumption, it was done until the point at which the assessment of further papers was not likely to generate more fundamental insights, in accordance with the principle of theoretical saturation from grounded theory (Flick, 1991; Strauss, 1998). This point was reached when a total of 28 recent articles that deal with one of those forms of consumption were reviewed. The results are displayed in table 1.

Authors generally agree that "green consumption" only refers to environmental aspects (Amatulli et al., 2019; Johnstone & Hooper, 2016; Nguyen et al., 2019; Perera et al., 2018). Therefore, green consumption is a subpart of sustainable consumption.

Another term is "responsible consumption". It is sometimes specified as "socially responsible consumption" (Robinot et al., 2017; Shobeiri et al., 2016) or "environmentally responsible consumption" (Gupta & Agrawal, 2018). However, that does not necessarily mean, that only the respective aspects are covered. Xu et al. (2019) deal with socially responsible consumption but also include environmental aspects. Authors that use the term responsible consumption without further specification mostly only consider social aspects (Antonetti & Maklan, 2015; Rasool et al., 2019) but some include environmental aspects as well (Kumar & Chamola, 2019). All in all, it can be concluded that responsible consumption rather focuses on social issues and is therefore a subpart of sustainable consumption. Furthermore, it is noticeable that the term in its unspecific form is used in articles that are not published in top tier journals (Kumar & Chamola, 2019; Rasool et al., 2019; Xu et al., 2019) especially if its unspecific form is used.

Numerous authors deal with "ethical consumption". Most of them attribute social and environmental issues to it (Carrington et al., 2016; Gummerus et al., 2017; Jung et al., 2016; H. Lee & Cheon, 2018; Shaw et al., 2016; Wiederhold & Martinez, 2018; Williams et al., 2015). This indicates that sustainable consumption (as defined earlier) and ethical consumption are equivalents regarding the covered aspects. Only one of the reviewed articles about ethical consumption deals with social aspects only (Govind et al., 2019).

Figure 3 graphically displays the relations of the different forms of consumption discussed in this chapter.

2.1.3. German Millennials

As mentioned in the first part, consumers of different nations and therefore different cultural backgrounds show considerable differences in consumer attitude and behavior (Gentina et al., 2014; Gómez-Luciano et al., 2019; Kacen & Lee, 2002; Lapierre & Rozendaal, 2018; Manrai, 2018). This thesis focuses on consumers that are born and raised in Germany.

Literature agrees regarding the general generational groups but does not really agree on the exact division regarding the year of birth (Wey Smola & Sutton, 2002). An exemplary summary of different divisions is displayed in table 2[4].

Table 2 shows that the different divisions do not correspond 100% but are all very similar. For this thesis, the division of the Center of Generational Kinetics is used, as it is a specialized research institute that focuses on generational research and should therefore provide valid information.

Table 3 provides data of the German population divided into generational groups as defined by the Center for Generational Kinetics (2016). As children under ten years are assumed to play a minor role in purchasing decisions, this age group was left out. As well, people over 73 years were not considered as they only represent a small part of the population and will not be of major influence in the future.

As mentioned before, Millennials have a more eco-friendly attitude than older generations (Bedard & Tolmie, 2018) as they are the first consumers to grow up in a globally interdependent world (Bucic et al., 2012). Their presence in the workforce and their spending power is big (The Nielson Company, 2015), which is supported by the fact that they represent the second biggest generational group in Germany. Only the group of the Baby Boomers include more people, but their relevance fades due to their advancing age. As a result, German Millennials are all people that are born between 1977 and 1995 in Germany and were also raised in Germany.

2.1.4. Sustainable Food

There is a huge amount of different food products in the German market and some products are considered more sustainable than others. Table 4 provides an overview of the emitted GHG in gram per kg of the respective product.

These data allow to compare different products regarding their GHG emissions. However, emissions are not the only impact on the environment. The water consumption during the production process is as well important (BMU, 2016). Referring to the social aspect of sustainability, fair-traded products are said to be more sustainable than products from exploitation. This is especially important for imported products like bananas, cacao, or tea (BMU, 2016; Tegut, 2015). Regional products are said to be more sustainable than imported ones, as there are additional emissions as a result of transportation. This may not always be the case, for example, oversea ships can be more C02 efficient than truck transportation. Small decentralized processing structures also tend to be less energy efficient than large companies (Penker & Elmar, 2015). This illustrates how multi-layered and non-transparent sustainability is and that regionally produced products are not necessarily more sustainable than others. Even if that is not generalizable, it may

information regarding the research on sustainable consumption.

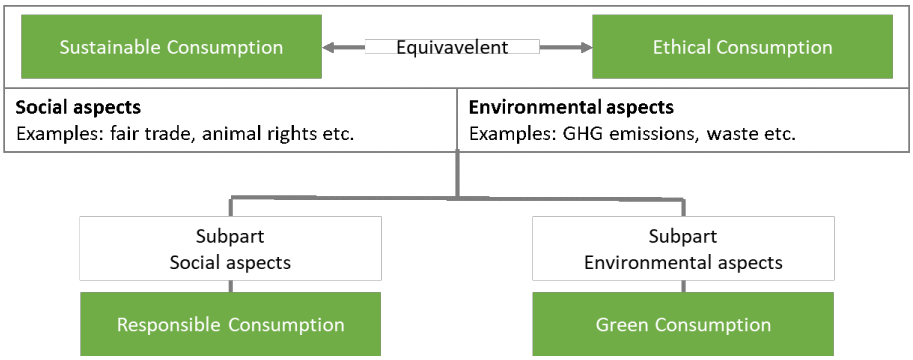
[3]Journal rank was obtained from Scimago Journal and Country Rank (https://www.scimagojr.com/) on 11.12.2019.

[4]See the appendix section A1.4 for the research strategy and additional information regarding German Millennials.

Table 1: Literature review of sustainable consumption and related forms of consumption

Kind of consumption	Covered aspects	References	Journal	Journal rank[3]
Sustainable	Environmental, Social	Prothero et al., 2011	Journal of public policy & marketing	Q1
Sustainable	Environmental	Chen et al., 2019	Journal of marketing research	Q1
Sustainable	Environmental, Social	Salciuviene, Buenaventura, & Lee, 2019	Engineering economics	Q2
Sustainable	Environmental, Social	Minton, Jeffrey Xie, Gurel-Atay, & Kahle, 2018	International journal of consumer studies	Q2
Sustainable	Environmental, Social	Scott & Weaver, 2018	Journal of public policy & marketing	Q1
Sustainable	Environmental	Brough et al., 2016	Journal of consumer research	Q1
Sustainable	Environmental, Social	Luchs et al., 2015	Journal of marketing management	Q1
Green	Environmental	Nguyen, Nguyen, & Hoang, 2019	Sustainable development	Q1
Green	Environmental	Amatulli, De Angelis, Peluso, Soscia, & Guido, 2019	Journal of business ethics	Q1
Green	Environmental	Perera, Auger, & Klein, 2018	Journal of business ethics	Q1
Green	Environmental	Johnstone & Hooper, 2016	Journal of marketing management	Q1
Responsible	Environmental, Social	Kumar & Chamola, 2019	Serbian journal of management	Q3
Responsible	Social	Rasool, Shakur, ani Mohammad, Mughal, & Awang, 2019	International journal of business & society	Q3
Socially Responsible	Social	Shobeiri, Rajaobelina, Durif, & Boivin, 2016	International journal of market research	Q1
Socially Responsible	Social	Robinot, Ertz, & Durif, 2017	International journal of consumer studies	Q2
Socially Responsible	Social	Song & Kim, 2018	Journal of business ethics	Q1
Socially Responsible	Environmental, Social	Xu, You, & Liu, 2019	Social behavior and personality	Q3
Environmentally Responsible	Environmental	Gupta & Agrawal, 2018	Corporate social responsibility and environmental management	Q1
Responsible	Social	Antonetti & Maklan, 2015	International journal of market research	Q1
Ethical	Environmental, Social	Hiller & Woodall, 2019	Journal of business ethics	Q1
Ethical	Social	Govind, Singh, Garg, & D'Silva, 2019	Journal of business ethics	Q1
Ethical	Environmental, Social	H. Lee & Cheon, 2018	Journal of international consumer marketing	Q2
Ethical	Environmental, Social	Gummerus, Liljander, & Sihlman, 2017	Journal of business ethics	Q1
Ethical	Environmental, Social	Carrington, Zwick, & Neville, 2016	Marketing theory	Q1
Ethical	Environmental, Social	Shaw, McMaster, & Newholm, 2016	Journal of business ethics	Q1
Ethical	Environmental, Social	Williams, Germov, Fuller, & Freij, 2015	Appetite	Q1
Ethical	Environmental, Social	Jung, Kim, & Oh, 2016	Journal of business ethics	Q1
Ethical	Environmental, Social	Wiederhold & Martinez, 2018	International journal of consumer studies	Q2

Own representation



Own representation

Figure 3: Different forms of consumption

Table 2: Different divisions of generations depending on age

Reference	Baby Boomers	Generation X	Generation Y - Millennials	Generation Z
McCrindle and Wolfinger (2014)	1946 - 1964	1965 - 1979	1980 - 1994	1995 - 2009
Brühl (n.d.)	-1964	1965 - 1975	1976 - 1998	1999 -
Meredith and Schewe (1994)	1946 - 1964	1965 - 1979	1980 - 2000	
Center for Generational Kinetics (2016)	1946 – 1964	1965 – 1976	1977 – 1995	1997 - TBD

Own representation

Table 3: Generations overview

	Baby Boomers	Generation X	Generation Y - Millennials	Generation Z
Birth range	1946 - 1964	1965 – 1976	1977– 1995	1996 - TBD
Age	73 – 55 years	54 – 43 years	24 – 42 years	23 – 10 years
Population in Germany	21.151.208	14.354.981	19.721.652	11.378.281

Own representation based on Statistisches Bundesamt (2019) and Center for Generational Kinetics (2016)

be true for a majority of cases (eKitchen, 2017; Moderne-Landwirtschaft, 2016). Therefore, this thesis will assume that regional products have a lower impact on transportation than others. Another important aspect is the packaging of the food, which is also multi-layered. A clear and general distinction of packaging in sustainable and not sustainable is not possible[5], as various factors like packaging material, the durability of the goods, and others have to be taken into account (Emblem, 2012; Sonneveld, James, Fitzpatrick, & Lewis, 2005).

To sum it up, the ecological aspect of a food product's sustainability depends on several different factors. These can be subordinated to three main categories: the impact during production, the impact from transportation and packaging. Still, it is not sensible to try to generally distinguish the products in sustainable and unsustainable and this is also not the goal of this thesis. Instead of following a black and white logic, this thesis focuses more on relative sustainability. For example, table 4 allows to assume that meat has a higher impact from production than vegetables. The table is displayed in an ascending order regarding their emissions (conventional agriculture). If the products are simply split into three approximately even groups, it results in the limit values regarding the C02 emissions as displayed in table 5.

Also, as previously described, the impact from transportation is lower if a product was produced regionally, than if it was produced at the other end of the country or even imported from another place in the world. Regarding packaging, it is impossible to determine which kind of packaging is more or less sustainable as extensive research showed[6]. The only sensible evaluation here is to distinguish between packed and unpacked products.

Furthermore, sustainable consumption also includes social aspects as described in chapter 2.1.2. In the further process of this thesis, the focus hereby lies on fair-trade and compliance with animal rights, as derived from the definition of sustainable consumption in chapter 2.1.2.

This thesis is especially aiming at exploring the attitude, thought processes, and beliefs regarding sustainable behav-

[5] See the appendix section A1.4 for a detailed research regarding sustainable packaging.

[6] See the appendix section A1.4 for a detailed research regarding sustainable packaging.

Table 4: Greenhouse gas emissions of the respective product

Product	Conventional agriculture C02 equivalent emission in g/kg	Ecological agriculture C02 equivalent emission in g/kg
Vegetables fresh	153	130
potatoes fresh	199	138
Tomatoes	339	228
Vegetables frozen	415	378
Vegetables canned	511	479
White bread	661	553
Mixed bread	768	653
Pastry	919	770
Baked goods	938	838
Milk	940	883
Yoghurt	1,231	1,159
Beef	1,311	11,374
Curd/cream cheese	1,929	1,804
Eggs	1,931	1.542
Pork	3,252	3,039
Poultry	3,508	3,039
Potatoes dry	3,776	3,354
Pork frozen	4,282	4,069
Poultry frozen	4,538	4,069
Pommes-frites	5,728	5,568
Cream	7,631	7,106
Cheese	8.512	7.951
Beef frozen	14.341	12.402

Own representation based on (BMU, 2016)

Table 5: Impact dimensions and evaluation sheet

	Low impact (see table 4)	Medium impact	High impact
Impact from production	Up to 900 C02 equivalent emission in g/kg	900 to 2,000 C02 equivalent emission in g/kg	Over 2,000 C02 equivalent emission in g/kg
Impact from transportation	Regional product	German product	Imported product
Impact from packaging	Unpacked product		Packed product

Own representation

ior. Therefore, it is not absolutely necessary to develop a valid and generalizable quantitative measure sustainability, but more what consumers perceive as sustainability and for example what happens in their heads when choosing a product, they perceive as less sustainable as available alternatives. Therefore, in the process of this work, “sustainable product" or “unsustainable products” respectively products that are “more” or “less” sustainable than others, refer to the peoples’ perception of the sustainability of these products and not to the evaluation sheet developed in this chapter. The developed sheet will only be used in the later process of this thesis in order to evaluate the actual purchase behavior of the participants.

2.2. Consumer Behavior

In the following section, the Theory of Planned Behavior (TPB) is described as the guiding theoretical framework for this thesis explaining the behavior of consumers.

The TPB is said to be the most popular theory to explain human behavior (Astrøm & Rise, 2001; Dean, Raats, & Shepherd, 2012). It has been successfully applied in various different areas of behavior, ranging from health behavior (Astrøm & Rise, 2001) to career choices (Moore & Burrus, 2019). More relevant for this thesis is its applicability for purchasing behavior (Ramayah, Lee, & Lim, 2012), sustainable behavior (Kaiser, Wölfing, & Fuhrer, 1999) and sustainable consumption (Hameed, Waris, & ul Haq, 2019; Robinot

et al., 2017; Sparks & Shepherd, 1992) and most important food consumption (Carfora et al., 2019; Dean et al., 2012; Schmidt, 2019; Shepherd & Raats, 1996; Shin, Im, Jung, & Severt, 2018; Sparks, Hedderley, & Shepherd, 1992; Testa, Sarti, & Frey, 2019). Especially when examining the relationship between attitude and behavior the TPB is the most widely researched theory (Armitage & Conner, 2001). Therefore, the TPB is selected as frame for this thesis. The reviewed references that recommend respectively use the TPB are displayed below in table 6.

As the TPB was created as an extension of the Theory of Reasoned Action (TRA) (Ajzen, 1991), both theories will be explained in the following chapters. The section will be concluded by a chapter examining diverging observations that emerged from prior literature having used the TRA and the TPB to explain consumer behavior.

2.2.1. The Theory of Reasoned Action

The development of the TRA was started by Martin Fishbein in the late 1960s. In previous decades, psychology failed to provide consistent findings about the correlation between attitude and behavior (Rossmann, 2011). Fishbein (1967) tried to solve that problem by (1) considering only a limited set of variables as determinants for behavior and (2) review of the relationships between determinant and traditional attitudinal measures (Rossmann, 2011). In that context, Fishbein (1967) formulated a goal and premises for a theory he called "theory of behavioral prediction" which became the TRA later (Rossmann, 2011).

Introduction to the Theory of Reasoned Action

The following subchapter is intended to give a broad overview of the TRA, its basic premises, and the general logic of the theory.

Basic Premises

There are two basic premises that have to be given for the TRA to be valid (Fishbein & Ajzen, 1980):

1. A rational thinking and acting human, that anticipates the meaning of his/her actions in consideration of all available information before deciding to behave in a certain way
2. The behavior to be evaluated is under deliberate control of the respective human

In fact, if people do not deal at all with their intentions and consequences of their behavior (e.g. automated behavior, manifested habits, addictive or aggressive behavior) or if they are not able to put their behavior into practice, the TRA reaches its limits (Fishbein & Ajzen, 1980; Rossmann, 2011).

Basic Model

If the two premises are given, the behavior can be precisely prognosed from the intention. That means, that people do what they intend to do. The intention is dependent on two determinants, the attitude towards the respective behavior and the social pressure or subjective norm that a person feels (Fishbein & Ajzen, 1980).

The relative importance of attitude and subjective norm depends in part on the intention under investigation. For some intentions, attitudinal considerations may be more important than normative considerations, for other intentions it may be vice versa. Additionally, the importance of the two determinants of intention can vary from person to person (Fishbein & Ajzen, 1980).

According to Fishbein and Ajzen (1980) attitudes are a function of beliefs. In simplified terms that means, a person who believes that performing a given behavior will lead to mostly positive outcomes will hold a favorable attitude toward performing the behavior and a person believing that the behavior will result in negative outcomes will hold an unfavorable attitude. These underlying beliefs are called behavioral beliefs.

Like attitudes, subjective norms are functions of beliefs as well, but a different form of beliefs. Rather than beliefs about the outcome of the behavior, these so-called normative beliefs are about whether specific individuals or groups think he/she should perform the behavior. In simplified terms that means, a person who believes that most people with whom he wants to comply, think he should perform a given behavior will lead him to feel the social pressure to do so. Again, it is vice versa if the person beliefs that the respective people think he should not perform the behavior (Fishbein & Ajzen, 1980).

Figure 4 summarizes the principle of the theory, how the behavior of a person can be traced back to that person's beliefs.

The following subchapters will each deal with one of the components (behavior, intention, attitude, subjective norm) of the theory, explaining them in detail, starting with the behavior, and then working through the model as displayed in figure 4 from right to left.

Behavior

Single Actions and Behavioral Categories

Fishbein and Ajzen (1980) state that it is important to distinguish single actions from behavioral categories. A single act is a specific behavior performed by an individual. To be able to measure a single action, it must be defined clearly enough to determine whether it has been performed or not. That may be easy for some actions like e.g. if attending church. For other actions however it may be very hard to determine if that person did really listen to what the priest said (Fishbein & Ajzen, 1980).

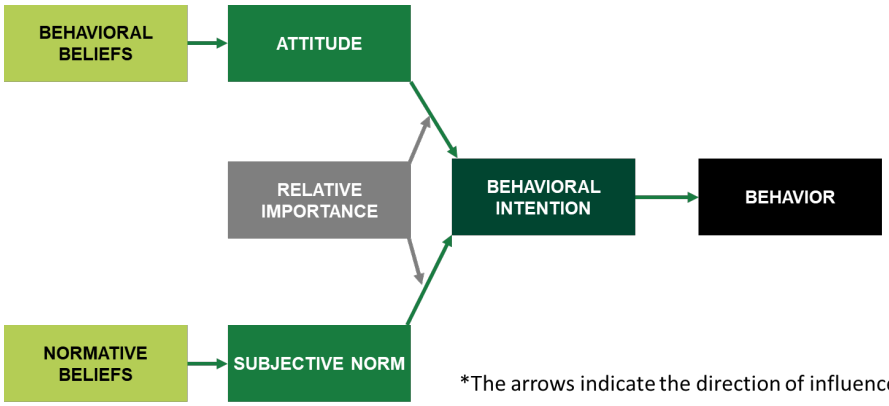
A behavioral category involves more than one single action. It can refer to a relatively narrow range of actions, like e.g. dieting or exercising or to a broader range like health maintenance or recreational activity. Behavioral categories cannot be observed directly. Instead, they are inferred from single actions, that are assumed to be instances of the general behavioral category (Fishbein & Ajzen, 1980). So e.g. exercising is inferred from the single actions of running, doing

[7] Journal rank was obtained from Scimago Journal and Country Rank (https://www.scimagojr.com/) on 24.01.2020.

Table 6: Overview of articles using the Theory of Planned Behavior

Type of behavior	Reference	Journal	Journal Ranking[7]
General behavior	Armitage and Conner (2001)	British journal of social psychology	Q1
Sustainable behavior	Kaiser et al. (1999)	Journal of environmental psychology	Q1
Ethical behavior	Wiederhold and Martinez (2018)	International journal of consumer studies	Q2
Socially responsible behavior	Han and Stoel (2017)	Journal of international consumer marketing	Q2
Internet usage	Rana, Slade, Kitching, and Dwivedi (2019)	Computers in human behavior	Q1
Health behavior	Astrøm and Rise (2001)	Psychology and health	Q1
Gambling behavior	Procter, Angus, Blaszczynski, and Gainsbury (2019)	Addictive behaviors	Q1
Career behavior	Moore and Burrus (2019)	The career development quarterly	Q2
General consumption	Kim, Ham, Yang, and Choi (2013)	International journal of hospitality management	Q1
General consumption	Johnstone and Tan (2015)	Journal of business ethics	Q1
General Consumption	Ramayah et al. (2012)	Journal of environmental management	Q1
Sustainable consumption	Hameed et al. (2019)	Environmental science and pollution research	Q1
Sustainable consumption	Robinot et al. (2017)	International journal of consumer studies	Q2
Green consumption	Sparks et al. (1992)	Social psychology quarterly	Q1
Food consumption	Sparks and Shepherd (1992)	European journal of social psychology	Q1
Food consumption	Dean et al. (2012)	Journal of applied social psychology	Q2
Food consumption	Schmidt (2019)	Food quality and preference	Q1
Food Consumption	Carfora et al. (2019)	Food quality and preference	Q1
Food Consumption	Testa et al. (2019)	Business strategy and the environment	Q1
Food Consumption	Shin et al. (2018)	International journal of hospitality management	Q1
Coffee consumption	Van der Merwe and Maree (2016)	International journal of consumer studies	Q2
Wine consumption	Maksan, Kovačić, and Cerjak (2019)	Appetite	Q1

Own representation



Own representation based on Fishbein and Ajzen (1980)

Figure 4: Factors determining a person's behavior

crunches, and doing push-ups.

Because a general behavioral category consists of many different single actions, the observation of only one or two of them will not provide an adequate measure of the respective category. On the one hand, this is because the observation of one or two single actions is simply a sample, which is too small, but on the other hand, the particular actions alone that are selected may not be valid indicators of the intended cat-

egory. For example, it is not clear if drinking coffee without sugar is a valid indication of dieting, as many people drink coffee without sugar, just because they like it that way. To be able to adequately measure a behavioral category it is necessary to observe a whole set of single actions and combine these measures to arrive at a general measure (Fishbein & Ajzen, 1980).

Behavioral Elements: Action, Target, Context and Time

Besides the action or actions themselves, it is important to consider the target at which a behavior is directed. The target is the aspect that is to be measured from a certain action e.g. if the particular person drinks beer or not. It can be also more specific like if this person drinks a certain brand of beer. These two examples demonstrate how different the measures can be depending on the target. As well the context is important, as it might make a difference if a person drinks beer at a bar or at home. Similarly, the time is important as it might also make a difference if someone drinks a beer in the evening or during the day (Fishbein & Ajzen, 1980).

Self-reported Behavior

Some behavior is not directly observable, which is why it might be necessary to rely on the actor's self-report. Usually, self-reports regarding behavior are quite accurate, but their accuracy cannot be taken for granted. There are no clear guidelines for the decision about the accuracy of a self-report, but they may be inadequate if there are strong reasons to suspect their inaccuracy. If self-reports can be used, they have some advantages over observations. Obviously, they require less effort, time, and money. However, it is not an adequate measure to ask a person directly whether she engaged in a behavioral category (e.g. dieting) as different people might have different understandings of the respective behavioral category. Here as well, a specific set of single actions has to be defined previously and the person of interest has to report his behavior regarding these specific actions. Similar to the procedure for direct observations, these data are then used to construct an index of the behavioral category (Fishbein & Ajzen, 1980).

Intention

Definition of Intention

The intention is defined as "a person's location on a subjective probability dimension involving a relation between him and some action" (Fishbein & Ajzen, 1975, n.p.), or in other words the behavioral intention is a measure of the likelihood that a person will engage in a certain behavior. According to the TRA, the intention is the immediate determinant of a person's behavior (Fishbein & Ajzen, 1980).

Determinants of Intention

As described before, Fishbein and Ajzen (1980) claim that behavioral intention is determined by (1) a personal attitude and (2) a social norm. In most instances, personal attitude and subjective norm are in agreement and the prediction of intention is relatively straightforward. Sometimes, however, both determinants may not be in agreement. If that is the case, the relative importance of the two determinants is highly relevant. This relative importance is dependent on the kind of behavior in general as well as the elements of behavior. A variation of action, target, context, or time might result in a different weighting of attitude and subjective norm. For example, for competitive behaviors, the attitude is more important and for cooperative behaviors, the subjective norm is more important (Fishbein & Ajzen, 1980).

Besides, Fishbein and Ajzen (1980) also indicate that individual or external factors influence the relative importance of the two determinants. These individual differences include demographic variables like age or sex, personality traits like introversion or extroversion, and many others. Ideally, the relative importance of attitude and the subjective norm would be available for each individual. It is important to notice that Fishbein and Ajzen (1980) state, that a person's own report does not provide satisfactory estimates. The relative importance can be estimated by performing a multiple regression analysis, the standardized regression coefficients can then be used as estimates.

Attitude

Definition of Attitude

The attitude is "a person's location on a bipolar evaluative or affective dimension with respect to some object, action, or event" (Fishbein & Ajzen, 1975, n.p.). In other words, this means that the attitude toward an object, action or event is simply a person's general feeling of favorableness or unfavorableness for that object, action, or event (Fishbein & Ajzen, 1980). It is important to mention that the TRA only refers to attitudes regarding behavior. It does not include traditional attitudes toward e.g. people or institutions (Fishbein & Ajzen, 1980). To be more accurate, the attitude refers specifically to the person's own performance of the behavior rather than the performance in general. E.g. a woman might be generally in favor of using birth control pills but as she is trying to become pregnant she will not intend to use such pills (Fishbein & Ajzen, 1980).

Determinants of Attitude

The attitude toward a behavior is determined by beliefs about outcomes of that behavior also called behavioral beliefs. Those beliefs are formed by a person's experiences. They may be the result of direct observation, the acceptance of information from outside sources, or the self-generation through inference processes. Some of the beliefs may persist over time while others may change, and new beliefs form. A person may hold a variety of beliefs, but the actual attitude is formed by only a small number (five to nine) of them, the so-called salient beliefs. However, it is also possible for more than nine beliefs to be salient and therefore determine a person's attitude. In any case, it is a limited number of beliefs that are salient, but it is impossible to determine with any precision the exact number. As the salient beliefs are the most present ones in a person's mind, it can be assumed that the first beliefs that a person emits are the salient ones. If a person believes a behavior to result in positive outcomes, she will acquire a positive attitude toward the behavior and vice versa (Fishbein & Ajzen, 1980). To sum up, the attitude towards a behavior is acquired automatically and simultane-

ously based on the salient beliefs (Fishbein & Ajzen, 1975).

Subjective Norm

Definition of the Subjective Norm

The subjective norm deals with the influence of the social environment on intentions and behavior. In easier terms, it means a person's perception that most people who are important to her think she should or should not perform a certain behavior. The more a person perceives that these other people think she should perform a behavior, the more she will intend to do so, and vice versa (Fishbein & Ajzen, 1980).

Determinants of the Subjective Norm

The subjective norm is determined by normative beliefs. As described previously, that does not mean what the person herself thinks about the behavior in question but what she believes her social environment opinion is. Of course, not everybody that is in contact with the person is important, but only some salient referents. Eliciting these salient referents is the first step in assessing the subjective norm.

Additionally, the general motivation to comply is important, as a person might weight the opinions of her referents differently. For example, regarding the previously mentioned use of birth control pills, a person will probably have a higher motivation to comply with her husband than with her friend.

In sum, subjective norms are based on the set of salient normative beliefs, each weighted by the motivation to comply (Fishbein & Ajzen, 1980).

The Complete Model of the Theory of Reasoned Action

The behavior of a person can be explained and predicted from her behavioral intention. This intention in turn is determined by the attitude and subjective norm. These two factors might be weighted differently which plays an important role in determining the intention. This relative importance can be influenced by external factors like demographic variables (e.g. age or sex), attitudes (not toward the behavior in question but related people or institutions), or individual personality traits.

The attitude is determined by salient beliefs about the outcomes of the behavior and how the person evaluates these outcomes as well as how likely she thinks that outcome is (belief strength). The outcome evaluation and the belief strength can as well be influenced by the previously mentioned external factors.

The subjective norm is comprised of the beliefs that the salient referents (the most important social contacts) think the person should do regarding the behavior in question and the general motivation to comply with these referents. These beliefs and motivations can also be influenced by the external factors. The complete model of the TRA is displayed in figure 5.

It is important to stress, that the described external factors do not necessarily influence the determinants. The relations specified within the theory, however (unbroken lines in figure 5), are always assumed to hold (Fishbein & Ajzen, 1980).

2.2.2. Theory of Planned Behavior

The TPB was introduced in 1985 by Ajzen (1985) as an extension of the TRA and soon empirically tested (Ajzen & Madden, 1986; Schifter & Ajzen, 1985). The theory is explicitly explained in a later work of Ajzen (1991) and compared to the related TRA by Madden, Ellen, and Ajzen (1992).

The TRA is limited as it requires people to have direct volitional control over the behavior in question. Therefore, it fails to explain behaviors for which this is not the case (Ajzen, 1991). The main difference between the TRA and the TPB is that the latter includes a third component that, additionally to subjective norm and attitude, influences behavioral intentions (Rossmann, 2011). This third component is called the Perceived Behavioral Control (PBC). By adding this component, the second limitation of the TRA is lifted so that the TBP can also handle behaviors that are not under complete control (Ajzen, 1988). Other than that, the TPB works the same way as the TRA (Rossmann, 2011).

The following sub-chapter explains in detail the perceived behavior control component. After that, an extension of the TPB is introduced until the chapter is concluded with a critical view on the TPB.

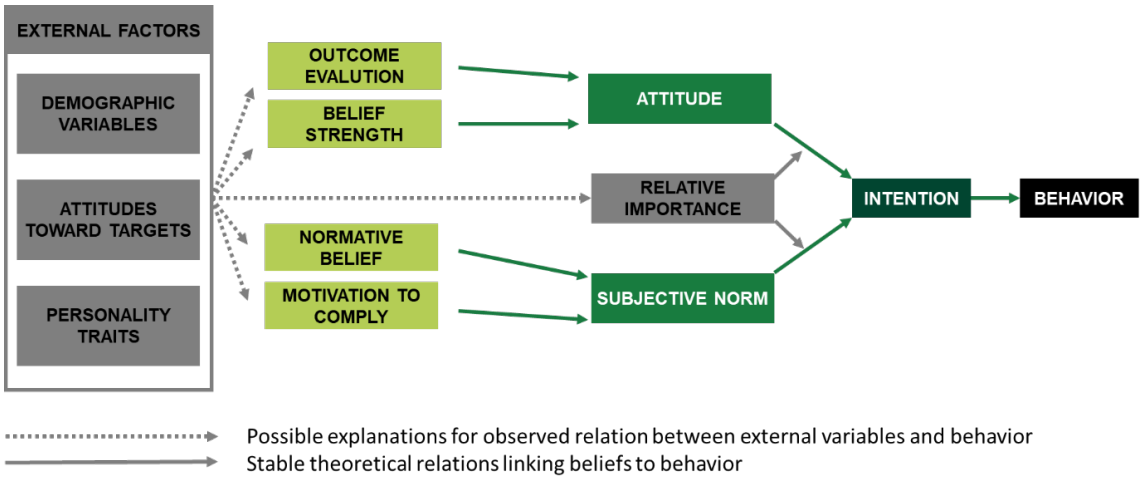
Perceived Behavioral Control

Definition of perceived behavior control

Many characteristics of an individual can influence the successful performance of a behavior. These include information, skills, and abilities but also the power of will, emotions, and compulsions. In addition, also external obstacles or circumstances can interfere with the performance of any behavior (Schifter & Ajzen, 1985). These factors collectively represent people's actual control or lack of control over the behavior (Ajzen, 1988). Most of these factors cannot or only partially be guessed in advance, let alone be measured. The individually perceived control of individuals over their actions is rather what these individuals think about their control than what it actually is. Whether a measure of perceived control can be used as a substitute for actual behavioral control is dependent on the accuracy of the perceptions. This accuracy is determined by the information a person has about the behavior, changes in requirements, or available resources, or the occurrence of new and unfamiliar elements (Ajzen, 1991). Basically, that means that if a person is familiar with a respective behavior e.g. because she has performed it numerous times before, the accuracy of her perception is higher respectively realistic. If the perception really is realistic, the PBC reflects the actual behavioral control quite well (Ajzen, 1985, 1988). Additionally, the PBC is quite easy to measure, as it can just be self-reported by the person in question (Rossmann, 2011).

Determinants of perceived behavior control

Similar to the attitude and the subjective norm, the PBC is also determined by beliefs, the so-called control beliefs (Ajzen, 1988). These are beliefs about what factors influence the actual execution of a behavior and how strong this influence is. These beliefs are mostly based on personal ex-



Own representation based on Fishbein and Ajzen (1980)

Figure 5: Complete model of the Theory of Reasoned Action

periences, experiences of the social environment, or other internal or external factors (Rossmann, 2011).

Also similar to the attitude and the subjective norm the salient beliefs are especially important in order to understand the PBC. For a single individual, it is likely that the first five to nine emitted factors are the salient ones (Ajzen, 1988). In a second step, these factors are evaluated regarding the probability that they will actually prevent the performance of the behavior in question (Ajzen & Madden, 1986).

Another approach is the direct assessment of the PBC without dealing with the detailed underlying beliefs. This can be done e.g. by asking the person about how she perceives her control about a certain behavior on just another 7-point-scale (Ajzen & Madden, 1986).

Influence of perceived behavior control

The PBC can influence the behavior in two different ways, either indirectly or directly (Madden et al., 1992). The indirect influence is based on the assumed influence on the behavioral intention. According to Ajzen (1988), this effect is not mediated by attitude or subjective norm. If a person thinks that they are not able to perform a given behavior, their intention to do so will likely be lower, even though subjective norm and attitude foster that behavior. E.g. the lacking use of Facebook among elder people is not necessarily a consequence of a negative attitude or a lacking social pressure, but simply their conviction not to be technically skilled enough for the use of such networks (Rossmann, 2011). Figure 6 provides a basic model of the TPB and indicates the influence of the PBC.

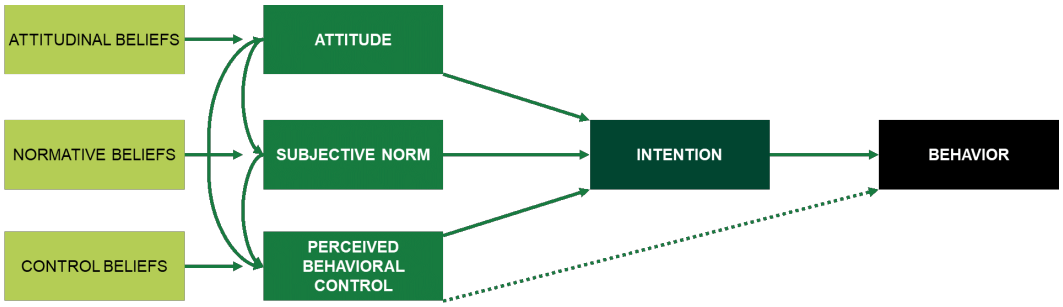
The direct influence reflects the actual and not the PBC in case a given behavior is in fact not controllable by a person. This influence only emerges if the PBC is realistic and does reflect the actual behavioral control (Ajzen, 1988). If e.g. a person wants to register to Facebook, but the internet connection is currently not working, this is a factor that cannot willingly be influenced by that person but has a direct impact on the performance of the behavior (Rossmann, 2011). In figure 6 this possible influence is displayed as a broken arrow.

Self-efficacy as Additional Determinant

Ajzen, 2002 states that the term "perceived behavioral control" could be misleading and should be interpreted as perceived control over the performance of behavior. Put in this way, it can be seen, that. PBC captures an individual's perception of their ability to perform a behavior, which is a product of the environment surrounding the individual. A related concept is the self-efficacy, that captures the perceived ability to perform the desired behavior (Bandura & Walters, 1977). Behavior and choices are also impacted by self-efficacy according to Bandura (1986). Self-efficacy and PBC can reflect both internal and external factors, and the extent to which they reflect one or the other is an empirical question. Parkinson, David, and Rundle-Thiele (2017) argue that self-efficacy primarily reflects internal factors, whereas PBC focuses rather on external factors. Previous studies have used the terms PBC and self-efficacy interchangeably (Bui, Droms, & Craciun, 2014). However, more research identifies them as separate constructs (Rhodes & Courneya, 2003; Yap & Lee, 2013).

Ajzen (1991) defined PBC as the individual perception of how easy or hard it is to perform a behavior, and the presence or absence of resources and opportunities to do it. However, other authors have argued that control over performing a behavior is different from how difficult people perceive the performance of the respective behavior (Chan & Fishbein, 1993; Sparks, Guthrie, & Shepherd, 1997). Therefore, PBC in this thesis is defined as the extent to which the performance of the behavior is up to the individual, describing external obstacles (Ajzen, 2002).

The perceived ease or difficulty of performing a behavior itself, given opportunity and resources is defined as self-efficacy, which is based on self-efficacy expectancy, which is the individual conviction that one is able to successfully per-



Own representation based on Ajzen, 1991

Figure 6: Basic model of the Theory of Planned Behavior

form a respective behavior and therefore describing internal obstacles (Bandura & Walters, 1977). Ajzen (2002) argues that carefully selected items should be used for both self-efficacy and PBC to ensure high-internal consistency. Therefore, self-efficacy will be added as an additional determinant of intention and behavior in the framework of this thesis. The model of a modified TBP that will be the theoretical foundation of this thesis is illustrated in figure 7.

To achieve better clarity, the arrows depicting the mutually influencing relations between the four determinants (Ajzen, 1991) were removed.

Status Quo and Critical View on the Theory of Planned Behavior

Despite being well established and applied countless times, the theory is claimed to have some weaknesses as well. Its biggest one is probably the fact that it assumes behavior to be rational and conscious. This neglects spontaneous behavior or strong emotional influence (Jones, 1996; Rossmann, 2011).

Furthermore, some critics question the completeness of the components as past behavior is not included, which means that habitualization respectively repeated behavior is neglected (Jones, 1996; Klöckner & Blöbaum, 2010; Rossmann, 2011). Ajzen (2005) himself states that the TPB is also valid for habitualized behavior, as it was once triggered by a conscious decision and therefore relies on the same cognition as not-habitualized behavior. Other psychologists like Gollwitzer (1999) or Ouellette and Wood (1998) rather assume that habitualized behavior works without any cognitive participation. As findings on the applicability of TPB to habitual behavior are not consistent, the problem cannot be solved empirically at the moment (Rossmann, 2011).

Additionally, the meta-analytic review of Armitage and Conner (2001) states that the fact that the TPB often relies on self-reported behavior makes it vulnerable to self-presentational biases (Gaes, Kalle, & Tedeschi, 1978).

Jonas and Doll (1996) criticize the conceptualization of behavior as a singular act, as the TPB was originally designed to predict singular, observable behavioral acts. More complex behaviors including many single acts, behavioral outcomes, and goals cannot be explained by the TPB in the same manner respectively only with poorer validity.

According to Rossmann (2011), it is a further weakness of the theory that it does not clearly distinguish PBC and self-efficacy. However, this was already discussed and solved by the introduction of self-efficacy as the fourth determinant of intentions in the previous chapter.

Another weakness as described by Jonas and Doll (1996) and Wirth, Von Pape, and Karnowski (2008) is the fact that the interplay of the components of the TPB only represents a limited time frame, which does not allow to follow the processual character of the relation of attitude and behavior. That is why some authors like Ouellette and Wood (1998) demand the inclusion of past behavior as a further determinant. Other authors like Wirth et al. (2008) suggest the introduction of a feedback loop between behavior and the three determinants in order to display the processual character.

The most important critic is the fact that many studies using the TPB are not able to prove the causality of the relations. Often the studies are based on cross-sectional analyses that identify all behavioral determinants and behavior at the same time. Strictly speaking, this means that it is not possible to clearly prove which of the determinants the cause of the behavior is. Even a reverse order, according to which the model components are a consequence of the behavior and not vice-versa, is possible in some cases (Rossmann, 2011).

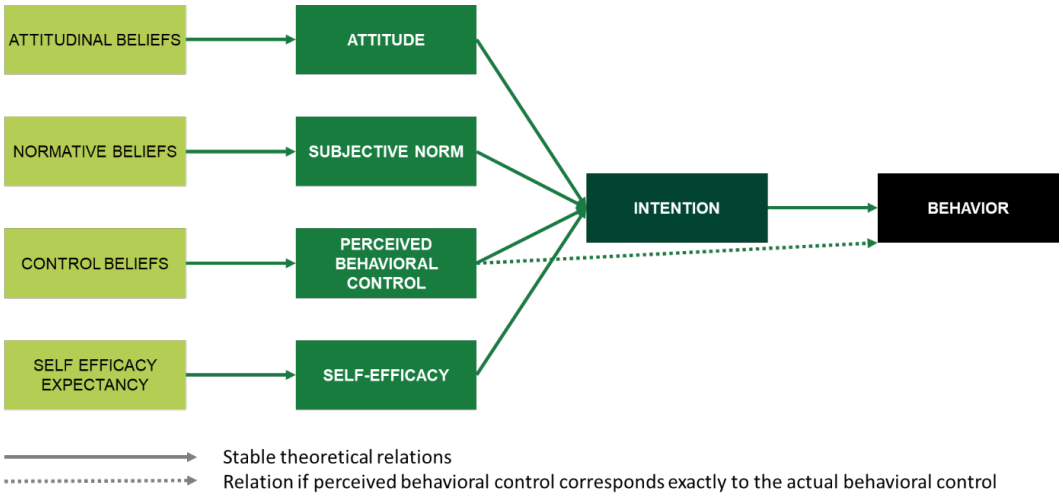
The further critic is mostly centered on the operationalization of the components like e.g. the elicitation of the salient beliefs (Rossmann, 2011).

Despite the weaknesses the TPB may have according to the previously mentioned authors, the TPB is among the most widely researched (Armitage & Conner, 2001) and empirically proven theories explaining behavior (Rossmann, 2011) and has still been frequently used in the last years as explained at the beginning of section 2.2. Many studies since the 1960s confirm its assumptions and various meta analyses confirm its empirical validity (Armitage & Conner, 2001; Han & Stoel, 2017; Rossmann, 2011).

The weaknesses of the TPB are displayed below in table 7.

2.2.3. Diverging Observations Regarding Consumer Behavior

As mentioned in the introduction, recent literature already picked up on discrepancies between certain determi-



Own representation based on (Ajzen, 1991, 2002; Bandura & Walters, 1977; Parkinson et al., 2017)

Figure 7: Modified model of the Theory of Planned Behavior

Table 7: Overview of the weaknesses of the extended Theory of Planned Behavior

Weakness
Explaining irrational behavior
Explaining repeated / habitual behavior
Neglecting the processual character of the relation between attitude and behavior
Divergence between self-reported and actual behavior
Missing proof of causality of relations
Operationalization of determinants

Own representation

nants and the actual consumer behavior. Two forms of the discrepancies are mentioned frequently but must be distinguished. One form is describing a divergence between consumer attitude and actual behavior on the other form in describing a divergence between the intention and the actual behavior. In the following chapter, these two forms are defined and distinguished.

The Attitude Behavior Gap and Related Concepts

Attitude Behavior Gap

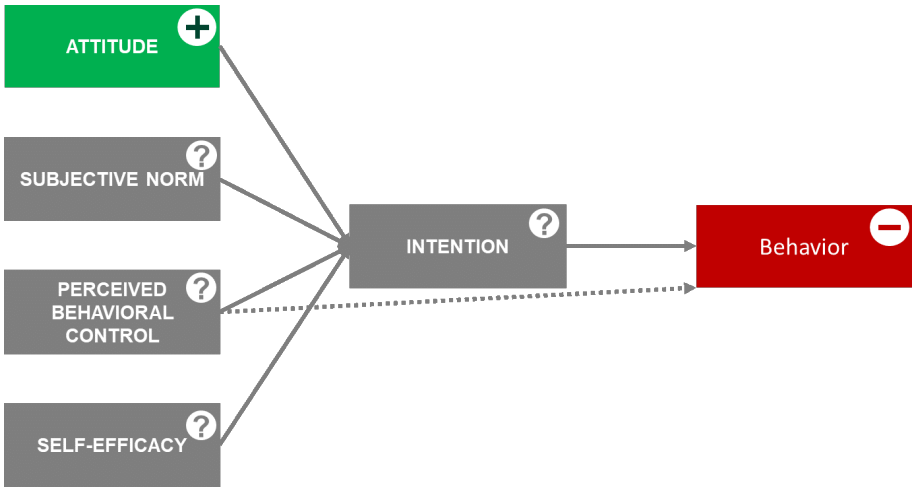
This thesis will subsume all terms describing the divergence between attitudes, values, concerns, opinions etc., and the actual behavior under the term ABG as derived from the TPB (Ajzen, 1991). This phenomenon is graphically shown in figure 8.

If attitude and behavior do not correspond, this divergence can be explained by the effect of an influence of the subjective norm, the PBC or the self-efficacy as it is not predefined what the intention is. For example, it could be possible that even though the attitude regarding a certain behavior is positive, the PBC and self-efficacy lead to a negative intention and a therefore a respective behavior, which is not corresponding to the attitude.

Other frequently mentioned terms describing the same phenomenon, are the value action gap (Ferguson et al., 2017) or the concern behavior gap (Wei et al., 2017).

Academic literature indicates the existence of this gap for different industries and countries around the globe, like e.g. the wine market (Schäufele & Hamm, 2018) and the apparel industry (Wiederhold & Martinez, 2018) in Germany, the pork consumption in Brazil (de Barcellos, Krystallis, de Melo Saab, Kügler, & Grunert, 2011) or organic food consumption in Denmark (Aschemann-Witzel & Niebuhr Aagaard, 2014). Numerous other authors deal with the ABG, which is displayed in an overview in table 8.

The overall attitude in Germany regarding sustainability is positive according to a study done by the German federal office for environment. More than 60% of the Germans think protecting the environment is important (Umweltbundesamt, 2019b). As Millennials are said to have a more eco-friendly attitude than older generations (Bedard & Tolmie, 2018), it can be assumed that German Millennials also have a positive attitude regarding sustainability. On the other hand, the consumption of unsustainable food is not declining. The GHG emission due to food products was stable since the beginning of the millennial (Umweltbundesamt, 2020). Supporting that fact, the meat consumption in Germany has been stable during the last twenty years (Bundesanstalt für Land-



Own representation

Figure 8: The attitude behavior gap in the extended model of the Theory of Planned Behavior

wirtschaft und Ernährung, 2015; Ernährung, 2019) and the number of chickens kept in cages rose more than 30% in the last ten years (Statistisches Bundesamt, 2019). As these examples illustrate and the fact, that prior literature indicates the ABG for several other markets around the world and in Germany, it can be assumed that there is an ABG regarding the sustainable food consumption of German Millennials as well.

Intention Behavior Gap

In addition to the ABG, some authors also mention the intention behavior gap (Frank & Brock, 2018; Liu, Wang, & Koehler, 2019; Mack et al., 2019; Nguyen et al., 2019; Pfeffer, Englert, & Mueller-Alcazar, 2019) or the intention action gap (Gabler, Butler, & Adams, 2013; Holdershaw, Gendall, & Wright, 2011; Kersten, McCambridge, M. Kayes, Theadom, & McPherson, 2015; Kuo & Young, 2008; Saddawi-Konefka, Schumacher, Baker, Charnin, & Gollwitzer, 2016). The latter seems to apply especially in medical science. All gaps that describe a divergence between the intention and the behavior are labeled as intention behavior gap in the framework of this thesis. The attitude behavior gap and the intention behavior gap are displayed in figure 9.

The figure 9 compares the two gaps, which makes the difference between them obvious. The ABG allows the other determinants besides from the attitude to account for an intention that does not correspond to the attitude and therefore explain the ABG. The intention behavior gap neglects the determinants of intention as it presumes it to be, for example, positive. Thus, the determinants cannot provide an explanation of the gap within the framework of the TPB and requires seeking explanations outside of the theory's logic.

The intention behavior gap can be a special case of the ABG if the attitude corresponds with the intention, but these two do not correspond with the behavior. In order to investigate the problem in a more holistic way, this thesis will investigate the ABG covering all the determinants of behavior.

Findings on the Attitude Behavior Gap

Some prior literature also aimed at qualitatively elaborating explanations for the ABG. Aschemann-Witzel and Niebuhr Aagaard (2014) examine the gap in the context of organic food products in for young consumers in Denmark and Wiederhold and Martinez (2018) look into the German green apparel industry. Explanation approaches include the price premium people cannot or do not want to pay (Aschemann-Witzel & Niebuhr Aagaard, 2014; Wiederhold & Martinez, 2018) and the availability of respective products (Aschemann-Witzel & Niebuhr Aagaard, 2014; Wiederhold & Martinez, 2018). While Wiederhold and Martinez (2018) identified a lack of knowledge regarding sustainability as a barrier, Aschemann-Witzel and Niebuhr Aagaard (2014) explicitly stated that a lack of knowledge regarding organic products is no barrier for the young Danish consumers. Wiederhold and Martinez (2018) also stated the missing transparency, consumers' inertia, and habits as explanations, factors that could also apply to other products than clothing.

2.3. Summary

According to the TPB, consumer behavior is determined by the intention which is in turn defined by attitude, subjective norm, PBC (Ajzen, 1991), and self-efficacy (Ajzen, 1991, 2002; Bandura & Walters, 1977; Parkinson et al., 2017). Prior literature indicates a gap between the attitude and the actual behavior regarding sustainable consumption for different consumer groups and in different markets (K. Lee, 2008; Moraes et al., 2012; Nguyen et al., 2019; Wiederhold & Martinez, 2018; Young et al., 2010).

So far, the consumer behavior of German Millennials regarding sustainable food consumption has not been investigated despite the market's enormous economic (Bundesministerium für Wirtschaft und Energie, 2019), environmental (Grünberg et al., 2010; Osterburg et al., 2009; Umweltbundesamt, 2013) and social relevance (Flatley, 2019; Handelsblatt, 2013; PETA, 2019).

Table 8: Overview of different articles dealing with the attitude behavior gap

Gap	Focus	References	Journal	Journal rank
Attitude behavior gap	High-cost situations	Farjam, Nikolaychuk, and Bravo (2019)	Ecological economics	Q1
Attitude behavior gap	Usage of sustainable materials in private house building	Götze and Naderer (2019)	International journal of sustainable development & world ecology	Q1
Value action gap	Deforestation behaviors in Guatemala	Robb, Haggar, Lamboll, and Castellanos (2019)	Environmental conservation	Q1
Attitude behavior gap	Apparel industry in Germany	Wiederhold and Martinez (2018)	International journal of consumer studies	Q2
Attitude behavior gap	Wine consumption in Germany	Schäufele and Hamm (2018)	Food quality and preference	Q1
Attitude behavior gap	Social impact assessment	Grieco (2018)	Nonprofit management & leadership	Q1
Value action gap	Imitation of visible green Behavior	Babutsidze and Chai (2018)	Ecological economics	Q1
Concern behavior gap	Environmentally friendly behavior	Tam and Chan (2018)	Global environmental change	Q1
Attitude behavior gap	Environmental education	Redondo and Puelles (2017)	International research in geographical and environmental education	Q2
Attitude behavior gap	Responsible consumption	Antonetti and Maklan (2015)	International journal of market research	Q1
Attitude behavior gap	Food consumption in Denmark	Aschemann-Witzel and Niebuhr Aagaard (2014)	International journal of consumer studies	Q2
Attitude behavior gap	Sustainable tourism	Juvan and Dolnicar (2014)	Annals of tourism research	Q1
Attitude behavior gap	Consumption in the UK	Moraes et al. (2012)	Journal of marketing management	Q1
Attitude behavior gap	Pork consumption in Brazil	de Barcellos et al. (2011)	International journal of consumer studies	Q2
Attitude behavior gap	Consumption in the UK	Young et al. (2010)	Sustainable development	Q1
Value action gap	Sustainable energy	Flynn, Bellaby, and Ricci (2009)	Sociological review	Q1
Value action gap	Waste recycling in Hong Kong	Chung and Leung (2007)	Environmental management	Q1

Own representation

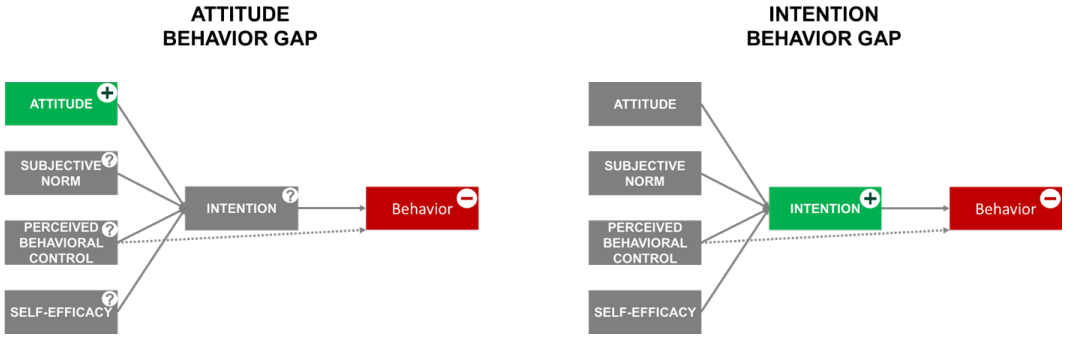
As described in the previous chapter 2.2.3 it can be assumed that there is an ABG regarding the sustainable food consumption of German Millennials as well.

As consumer behavior regarding various kinds of products is different (Pelau, 2011) it is sensible that also the reasons and explanations of the ABG regarding these kinds of products differ to the findings that prior works with similar research objects, like e.g. Wiederhold and Martinez (2018) elaborated. Thus, the following research question, as already stated in the introduction of this thesis, is to be examined:

How can the assumed attitude behavior gap regarding the sustainable food consumption of German Millennials be explained, respectively what are specific reasons for the gap?

3. Research Design

In order to investigate the research question, the assumed ABG must be deeply explored. Therefore, an exploratory qualitative approach was applied. It was especially important to explore specific detailed reasons that lead to a per-



Own representation

Figure 9: Comparison of the attitude behavior gap and the intention behavior gap

son's respective attitude, subjective norm, or PBC down to the basic beliefs that underly these different determinants according to the TPB (Ajzen, 1988; Fishbein & Ajzen, 1980). To expose these non-obvious and sometimes even unconscious beliefs an explorative, qualitative research design was necessary (Flick, 2018; Giri & Biswas, 2019). In order to get valid and comprehensive data, the approach was triangulated by combining different methods (Bohnsack, Meuser, & Geimer, 2018; Flick, 2018).

In a first step, the participants will be observed during a shopping trip in order to get insights about their behavior when shopping groceries. In a subsequent step, their actual purchases are examined before they are interviewed. The arrangement of these three data collection steps is sensible. If the interview was made before the observation, participants may be encouraged to think about their consumer behavior which would be likely to bias their behavior during the observation.

3.1. Recruitment and Sample

The participants were eleven consumers born between 1977 and 1995 and raised in Germany. They are intended to represent the whole group of German Millennials which is why a balanced spread regarding age and education was achieved. All participants were contacted by phone and agreed to take part in the study which cumulated in a participation rate of 100%. This might be due to the fact that all participants were first or second-grade social contacts of the researcher. They were given background information about the overall study goal and a meeting with each participant was arranged at the supermarket where they most frequently shop for their food.

Recruitment, observation, and interviews took place in January and February 2020. Observation and interviews were made on the same day. Before starting the data collection, each of the participants was asked to sign a declaration of consent regarding the usage of the data.[8]

Additional background information like the demographic data of the participants are displayed in table 9.

Participant 6 was a special case, as she brought her two-year-old daughter along for the shopping trip as she usually does. All other participants were alone with the researcher.

3.2. Direct Observation and Verbal Protocol Analysis

In a first step, the participants were observed during their regular grocery shopping in order to capture their actual consumer behavior as well as body language and external factors while buying food and beverages. According to the Verbal Protocol Analysis (VPA) (Ericsson, 2006; Ericsson & Simon, 1985), also known as the "think out loud" method (Reicks et al., 2003), participants were asked to share their thoughts and considerations in order to capture them immediately during the shopping. Using only retrospective methods like interviews may have lead to a loss of those thoughts and considerations.

3.2.1. Procedure

Orientation Interview

The researcher met the participants in front of the supermarket and gave them accurate and standardized instructions on the shopping protocol. After that, the researcher asked for circumstantial background information like the shopping list, the size of the household they are shopping for, or the available budget for today's shopping.

The Observation - VPA

After the orientation interview, the researcher walked with the participants who were asked to shop in their usual manner but, in addition, to verbalize all thoughts and considerations in the context of selecting the food and beverages. The researcher responded mostly to the participant by nonverbal signals to show that he was following the participant's thoughts and that he was interested, or by making a neutral comment. If the participant stopped speaking for a longer period (more than 60 seconds), the researcher used standardized non-leading sentences to encourage the monologue, such as "What are your thoughts?". In the standard VPA, the researcher does not respond to the participant's

[8] The declarations are available from the researcher on request.

[9] See appendix section A3.1 for additional information regarding the sample.

Table 9: Participants and additional information[9]

Proband No.	Gender	Birth year	Highest education	Current job	Household size	Urban(U) Rural (R)	Observation city	Observation shop
1	M	1993	Bachelor of Arts	Student	2	U	Munich Obergiesing	City Rewe
2	F	1978	Abitur	Foreign language correspondent	2	R	Pfaffenhofen	Lidl
3	F	1977	Vocational Training	Tax expert	2	R	Weissenhorn	Rewe
4	M	1991	Master of Eng.	Engineer	2	R	Senden	Lidl
5	M	1980	Vocational Training	Gas station owner	1	R	Pfaffenhofen	Rewe
6	F	1984	Vocational Training	Paralegal	4	R	Weissenhorn	Lidl
7	M	1994	Master of Science	Research Assistant	2	U	MunichSchwabing	Edeka Express
8	M	1990	Master of Science	Controller	2	U	Munich-Pasing	Lidl
9	M	1993	Bachelor of Arts	Student	1	U	Munich-Schwabing	Lidl
10	F	1984	Medical state examination	Physician	4	U	Munich Fürstenried	VollCorn Biomarkt

Own representation

questions or comments to avoid interrupting the process (Ericsson, 2006; Ericsson & Simon, 1985; Reicks et al., 2003). In an interactive, potentially busy, and loud environment, such as a supermarket, it is better to communicate with the accompanying person rather than to speak to oneself (Saarela et al., 2013).

Immediate Feedback Interview and Debriefing

Right after the shopping, short feedback interviews were conducted. The participants were asked general questions like "How did you feel about the shopping experience you just had? What was it like to think out loud?". The participants were also given an opportunity to ask questions or address concerns.[10]

3.2.2. Pretest

As pretest one observation was made prior to the actual data collection in order to test the applicability, completeness, comprehensibility, and quality of the research instrument (Raithel, 2008).

The pretest worked out fine. The participant stated he did not feel strange during the shopping trip, that he was instructed well and understood his task. As the participant did not keep the bill, reconstructing what he actually bought during the trip was rather time-consuming, as each product had to be recorded individually. As a result, the instruction given to the participant prior to the start of the observation was expanded, adding the request to keep the bill. Apart from that, the pretest worked out well and nothing else was changed. The pretest was used as regular observation, making the pretest participant "Participant 1".

3.2.3. Data Collection

As mentioned in section 3.1 data was collected from eleven participants over the course of about six weeks. One researcher accompanied each participant to her most frequently visited supermarket. Shopping trips took between seven and 25 minutes. Prior to the actual trip, additional information about the participants and the circumstances of the shopping were elicited with a short interview (see chapter 3.1.1).

The verbal protocol of each participant was recorded with a cable bound and smartphone assisted microphone that was clipped to the collar of the respective participants. Thus, the protocol of each shopping trip is saved as one audio file which was later transcribed verbatim. At the same time, the researcher noted all observed behavior of the participants (which products did she look at, which did she put back, what did she actually buy, body language, striking or unusual behavior etc.). Additionally, the researcher noted external factors like special offers, sold-out products, crowded areas in the store etc.

During the observation the audio gear failed three times, resulting in either a complete stop of the audio record (Partic-

[10] See the appendix section A2.1 for the full observation guide, including instructions and the debriefing guide.

ipants 6 and Participant 8) or in a record of incomprehensible noise (Participant 7). In these cases, no verbal protocol could be recorded. However, in these cases a protocol was written on the same day of the observation from the memory of the researcher who accompanied and observed the participant, thus still ensuring rich insights in the participants shopping behavior.

The transcript of the verbal protocol as well as the notes from the observation were combined into one consolidated and comprehensive report per participant. These reports are a combination of two transcript forms introduced by Mayring (2016). From the audio file transcript, which was the result of the verbal protocol, a summary transcription is created. Mayring (2016) suggests reducing the quantity of information in order to improve the quality content. Then, based on the concept of the commented transcript (Mayring, 2016), the notes from the observation are added, thus creating the consolidated report.

3.3. Analysis of Actual Purchase Decisions

In a subsequent step, the actual purchase decisions of the participants were evaluated by either analyzing the bill from the shopping trip that was observed and/or the shopping cart and/or the food supply stored at participants home or in order to see what the participant actually purchased. It is important to mention that this part of the study was not the main focus of this thesis, but rather an additional analysis, enabling a rough crosscheck of the self-reported attitude and behavior stated by the participants during the interviews and their actual purchased products. Thus, the result of this analysis should be interpreted with care, as results might not be universally valid or generalizable. While the observation and the VPA focuses on the overall behavior and thoughts during the shopping trip, this analysis focuses on the actually purchased products.

According to the impact dimensions of food products identified in chapter 2.1.4, the purchased goods are examined on a high-level. Therefore, a photo is taken of the bill, the shopping cart, and/or the participants food storage, if it was feasible. Based on the photos a list of the purchased products is created, and each product is evaluated regarding the three impact dimensions scheme elaborated in chapter 2.1.4. as well as regarding fair trade and organic production. The aspect of compliance with animal rights is excluded from the analysis as it was impossible to evaluate the compliance for the products after they got purchased, solely on the basis of the photos.

Impact from Production

The products were evaluated regarding their GHG impact according to the table 4, as far as they could be found in it. Then they were classified based on the three impact levels provided in table 5 in chapter 2.1.4 and received a score (low impact = 1, medium impact = 2, high impact = 3). By calculating the average score per participant, a comparable evaluation regarding the impact from production was achieved.

Impact from Transportation

To evaluate the purchase decisions regarding the impact from transportation, the share of percentages explicitly marked as regional products will be calculated for each participant.

Impact from Packaging

As described in chapter 2.1.4, an evaluation of sustainable packaging is complicated and cannot be done superficially. Therefore, in order to superficially evaluate the purchase decisions, the percentage of unpackaged products for each participant is calculated.

Fair Traded and Organic Products

Finally, the percentage of fair-traded products as well as a percentage of organic products out of all products is calculated.

3.4. Semi-structured Interviews

In a last step, the participants were interviewed in order to investigate their attitudes, subjective norm, and PBC in more detail, as well as to pick up on striking behaviors during the grocery shopping. By specifically asking for underlying beliefs forming attitude, subjective norm, and PBC these determinants can be explained and understood. To uncover these underlying, potentially hidden or unconscious beliefs, a semi-structured depth interview is a suitable method (Flick, 2018; Giri & Biswas, 2019).

3.4.1. Pretest

As pretest, one interview was done prior to the actual data collection in order to test the applicability, completeness, comprehensibility and quality of the research instrument (Raithel, 2008).

Again, the pretest was successful, and the participant gave a positive feedback. The interview generated a deep insight into the participant's attitude, subjective norm, PBC, and self-efficacy and provided some explanations for the ABG. However, it was often necessary to ask deeper questions and deviate from the key questions, as they mostly resulted in superficial answers. These deeper, detailed questions were noted and added to the interview guideline as examples. The pretest showed, that the interviewer needs to be flexible and adapt his questions if necessary, in order to unveil the hidden beliefs. However, the interview provided meaningful and purposeful insights, which is why it was used as a regular interview, so the pretest participant was "Participant 1".

3.4.2. Data Collection

Similar to the observation, the data was collected by one investigator that interviewed the previously mentioned participants according to the interview guideline[11]. The guideline was split into five parts. The first part was rather introductory with general questions like, "what do you think is sustainable food consumption?" or "how often do you go grocery shopping?" in order to get the participant to start talking

[11] See the appendix section A2.2 for the full interview guide.

about their general shopping behavior. The other parts were developed in accordance with the structure of determinants as described in section 2.2. The second part aimed at eliciting the attitude of the respective participant by questioning the attitude itself as well as underlying beliefs. The third part covered the subjective norm, similar to the previous part, covering the determinant itself by asking questions like "Which people from your social environment influence your choice in groceries?" as well as underlying beliefs. The fourth part is about the PBC and self-efficacy, again addressing the determinants themselves and underlying beliefs. The last part is optional in case the data collected up to that point are not rich enough in the eyes of the researcher. The question of that part center on the ABG and how it could be counteracted or anticipated.

The interviews were recorded with the same recording technique used for the observations and transcribed verbatim as suggested by Mayring (2016).

3.5. Qualitative Data Analysis

The analysis process follows a slightly different order than the data collection process. Here, data from the qualitative interviews were analyzed first in order to see how participants self-reported their behavior and of course attitude, subjective norm, PBC, and self-efficacy. In the following, the combined transcripts from the observation and VPA were analyzed in order to see how they actually behave during grocery shopping and what their actual thoughts are. In a last step, the actual purchase decisions are analyzed in order to match these against the self-reported behavior from the qualitative interviews.

The analysis of the interview transcripts and the combined reports of the observations and the VPA follows the principle of qualitative content analysis by Mayring (2010). With the intention to identify and summarize objectively as well as systematically the content of the gathered data and to draw conclusions from it (Mayring, 2010). This bears the advantages of a structured approach by maintaining openness (Kohlbacher, 2006). Basically, the analysis consists of three process steps. The first step is called summary and reduces the material into a manageable scope by preserving the essential content. Afterwards, the explication step helps to clarify the content and the last process step is structuring. This describes the most important part of the qualitative analysis by coding the text excerpts into suitable categories (Kohlbacher, 2006; Mayring, 2010). Qualitative content analysis is a proven and suitable tool for analyzing data within qualitative research projects. Besides the previously mentioned advantages, it enables the integration of context and different material as well as the use of quantitative steps for the analysis (Kohlbacher, 2006). To manage all the gathered data in this research project the software MAXQDA and Microsoft Excel are used to facilitate the content analysis.

3.6. Overview: Triangulated Research Process

The whole research process is graphically illustrated in figure 10.

That research design guarantees a multiple triangulation. By using different sources respectively questioning different persons at different times in different locations, a data triangulation is given, ensuring an enhanced gain in information (Bohnsack et al., 2018; Flick, 2011). Furthermore, a methodological (between-method) triangulation results from the usage of observations as well as interviews.

4. Findings and Interpretation

This part presents the findings. This means at first the results of the qualitative interviews are introduced, followed by the results of the observation and the VPA and the results from the analysis of the actual purchased products. The last section combines the findings from the previous sections in order to examine the occurrence of the ABG and develop explanation approaches.

4.1. Findings from the Qualitative Interviews

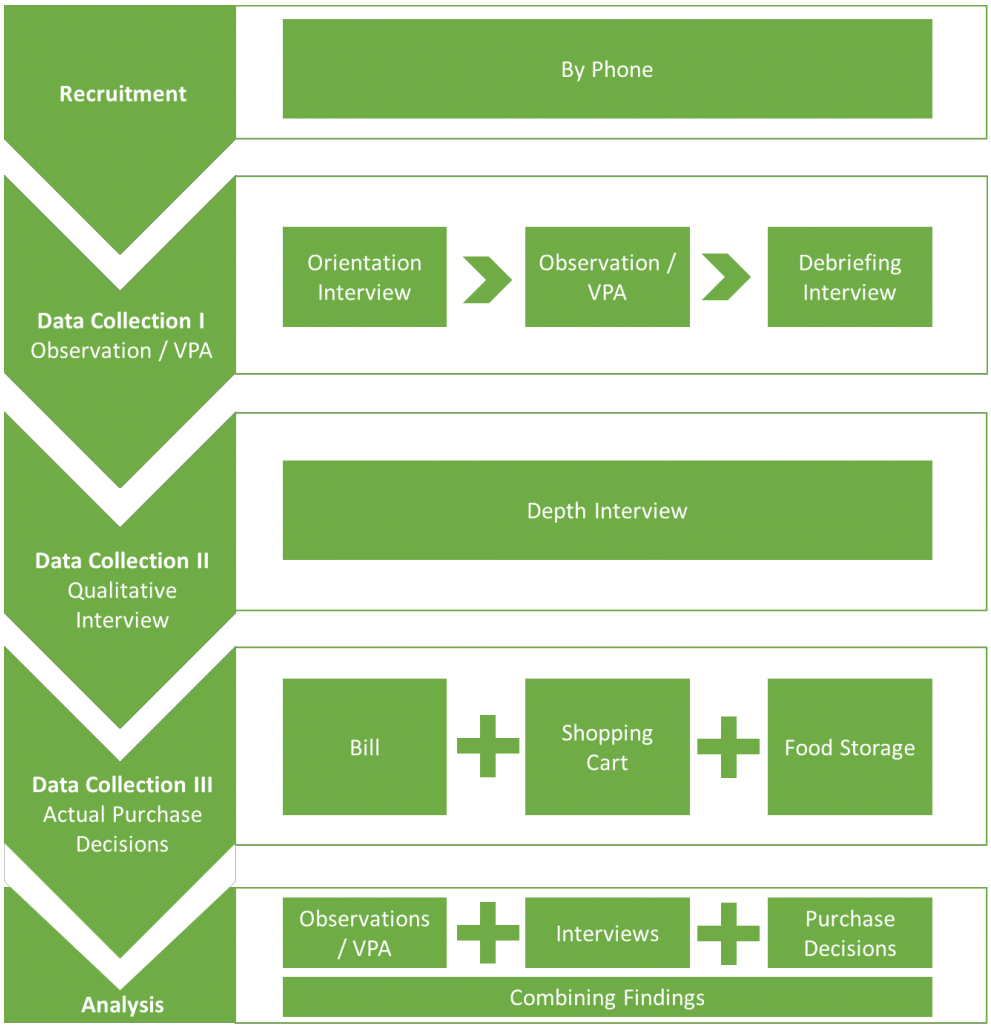
This section provides the findings of the qualitative interviews. At first, the general shopping behavior as self-reported by the participants is described. Subsequently, insights about the determinants of behavior according to the TPB are presented. The sources for the statements in that section are the interview transcripts with the respective participants. For example, the interview transcript of Participant 1 is shortened as "P1.I".

4.1.1. Findings from the General Shopping Behavior

Most of the participants stated that they do one big weekly grocery shopping trip and two additional smaller ones (P2.I; P3.I; P4.I; P6.I; P7.I; P10.I). The others indicated to do one (P3.I) or two (P1.I) trips per week. However, it seems like almost all participants do at least one bigger shopping trip per week.

Participant 1 and Participant 3 stated that they exclusively go grocery shopping in the evening after work when it is busy and sometimes some products are already sold out for the day (P1.I; P3.I). Participant 1 even said that this caused him to have a rather negative attitude regarding shopping in general and that he aims at getting over with the grocery shopping as fast as possible in order to only spend a minimum time in the store (P1.I). Participant 5 explained as well that he tries to minimize the time he is spending for grocery shopping (P5.I).

When it comes to choosing a store for their grocery shopping the biggest part of the participants named the proximity to their workplace respectively to their place of residence as one main driver (P1.I; P4.I; P5.I; P10.I). Other factors that were mentioned are the quality of the products (P4.I), the price compared to other stores (P9.I), the size of the product selection (P10.I) and the fact that some participants prefer to go to the same store every time because then they know exactly where to find which product, which increases shopping efficiency (P1.I; P5.I). This corresponds with the statements of Participant 1 and Participant 5 that they do not like grocery



Own representation

Figure 10: Triangulated research process

shopping and try to minimize the time they are spending for it, as previously mentioned (P1.I; P5.I).

The category scheme regarding general shopping behavior, providing an overview of the number of statements assignable to the respective category (n) as well as the total number of mentions (m) and the respective participants, is displayed below in table 10.

4.1.2. Attitude

Participants Evaluation of Attitude

It is striking that all participants, except for Participant 6, indicated to have a positive attitude regarding sustainable food consumption. While four participants (P1.I; P2.I; P3.I; P10.I) just stated a positive attitude, five participants (P5.I; P6.I; P7.I; P8.I; P9.I) added certain limitations to their positive attitude. For example, Participant 4 explained that even though sustainability is important for him, his goal is not even to consume fully sustainably but he rather tries to improve his consumer behavior only in certain areas (P4.I). Participant 5 added that he does not want to overdo it with thinking about sustainable consumption, as eating and drinking is meant to be an enjoyment (P5.I).

Some participants explicitly pointed out that their consumer behavior got more sustainable in the last two to three years (P1.I; P3.I; P7.I). Participant 7 explained in that context that his personal transition to sustainable consumption was nothing that happened overnight, but he rather described it as a process that started more than two years ago. He has been trying to improve his ecological footprint but does not really have the ambition to be completely sustainable today (P7.I).

The reason for this change in attitude could be the fact that many participants (P3.I; P4.I; P6.I; P7.I) feel affected by the media. As sustainability is gaining more and more attention, participants are made to think about the topic and reflect their consumer behavior, which seemingly benefits a

[12] See the appendix section A4.2 to get access to the full coding scheme.

Table 10: Category scheme regarding general shopping behavior[12]

Main Category	Sub-category 1	Sub-category 2	Mentions n	Mentions m	Participants
1 General shopping behavior	1.1 Reasons for the store choice	1.1.1 Large selection	1	1	P10.I
		1.1.2 Price	1	1	P9.I
		1.1.3 Proximity to work/place of residence	4	4	P1.I, P4.I, P5.I, P10.I
		1.1.4 Always the same shop (routine)	2	2	P1.I, P5.I
		1.1.5 Product quality	1	1	P4.I
	1.2 Negative attitude regarding shopping		2	6	P1.I, P5.I
	1.3 Shopping in the evening		2	2	P1.I, P3.I
	1.4 Number and kind of shopping trips	1.4.1 One per week	1	1	P3.I
		1.4.2 Two trips per week	1	1	P1.I
		1.4.3 One big trip plus two small trips per week	6	6	P2.I, P4.I, P6.I, P8.I, P9.I, P10.I
		1.4.4 Two - Three trips during the week, one big every two weeks	1	1	P7.I

Own representation

positive attitude regarding sustainable consumption.

Another explanation for an attitude change seems to be the birth of a child of one's own. The two participants, who are a parent, mentioned that the birth of their child had a great impact on their attitude (P6.I; P10.I). While Participant 10 stated the birth of her children had a positive influence on her attitude, as she wants to be a positive role model for her children regarding sustainable consumption (P10.I), Participant 6 explained that only with the birth of her children, she started thinking about what will happen to this world in the future and how she is influencing that (P6.I).

Even though most of the participants indicated to have a positive attitude, it is striking that half of the participants are currently not (yet) consuming as sustainable as they could and that they are aware of their potential for improvement.

Table 11 provides the category scheme regarding the participants' attitude evaluation, giving an overview of the number of statements assignable to the respective category (n) as well as the total number of mentions (m) and the respective participants.

[13] See the appendix section A4.2 to get access to the full coding scheme.

Underlying Beliefs

Understanding of Sustainable Consumption

When it comes to the underlying beliefs regarding the attitude, it is important to capture what people understand by sustainable food consumption. Therefore, the participants were asked to describe what sustainable food consumption means to them and what different aspects are covered by the term.

The most often mentioned aspect is the reduction of the consumption of products with excessive packaging especially if it is made out of plastic (P2.I; P3.I; P4.I; P5.I; P6.I; P7.I; P9.I; P10.I).

Almost as many participants mentioned the origin country of a food product as a determinant of sustainability. Regional products are perceived as more sustainable than products imported from foreign countries (P1.I; P2.I; P3.I; P4.I; P6.I; P7.I; P9.I; P10.I). The origin of products is related to another aspect that is frequently addressed by participants as an important part, namely the C02-footprint. Consuming products with the lowest possible C02 emissions during production is seen as sustainable (P3.I; P6.I; P7.I; P8.I; P9.I). As

Table 11: Category scheme regarding the attitude evaluation[13]

Main Category	Sub-category 1	Sub-category 2	Sub-category 3	Mentions		Participants
				n	m	
2 Attitude	2.1 Participants attitude evaluation	2.1.1 General	2.1.1.1 Positive attitude	5	5	P1.I, P2.I, P3.I, P10.I
			2.1.1.2 Limited positive attitude	5	8	P4.I, P5.I, P7.I, P8.I, P9.I
			2.1.1.3 Neutral attitude	1	1	P6.I
		2.1.2 Positive change in recent past		3	3	P1.I, P3.I, P7.I
		2.1.3 Change through own children		2	3	P6.I, P10.I
		2.1.4 Impact from media		4	6	P3.I, P4.I, P6.I, P7.I
		2.1.5 Aware of potential for improvement		5	5	P1.I, P2.I, P3.I, P7.I, P9.I

Own representation

the transport of products causes further C02 emissions it is only reasonable that imported products are considered less sustainable than regional products.

Also, participants frequently stated that reducing food waste is an important aspect of sustainable consumption (P1.I; P5.I; P6.I; P7.I; P9.I). This is connected to all the others as a reduction of food waste means fewer food products have to be manufactured, which is thereby influencing all other aspects.

Furthermore, it was striking that six out of ten participants stated that they had recently reduced (P2.I; P3.I; P4.I; P6.I; P7.I; P10.I) or even completely stopped their meat consumption (P1.I).

Additionally, many participants indicated to especially focus on consuming natural respectively organic products, partly as they believe organic products to be more sustainable than usual products and partly as they expect these products to be healthier than usual products (P1.I; P2.I; P4.I; P6.I; P7.I; P8.I; P9.I; P10.I)

Further issues that were only selectively mentioned by few participants are the decision to abstain from dairy products due to their high C02 emissions (P4.I), from products which required excessive quantity of water during production (P1.I; P7.I) and from products, which were produced under questionable circumstances regarding animal rights (P1.I; P8.I; P10.I). Some participants also claimed to specifically value fair traded products (P7.I; P8.I; P10.I).

The category scheme regarding the participants' understanding of sustainable food consumption, providing an overview of the number of statements assignable to the respective category (n) as well as the total number of mentions (m) and the respective participants, is displayed below in table 12.

Participants identified many different aspects of sustainable food consumption. However, some of them can be backtracked to one basic aspect. The origin of the products is important because of the emissions of GHG resulting from transportation. Also, meat and dairy products reduced diet is often also tied to the fact that there are high emissions during the production of these products (see chapter 2.1.4), which allows to subsume it under the aspect of reducing the GHG impact from production. As the reduction of impact from packaging was also an aspect that was frequently mentioned by participants this leads to the same three impact dimensions identified in chapter 2.1.4. Also mentioned in that chapter were the aspects of fair-traded products and the compliance with animal rights, both of which were also mentioned by the participants. Additionally, the reduction of food waste and the reduction of unhealthy / inorganic products were brought up, ultimately leading to the following seven main aspects of sustainability as described by the participants:

1. Reduction of the GHG impact from production
2. Reduction of the GHG impact from transportation
3. Reduction of the impact from packaging
4. Reduction of the consumption of not fair-traded products
5. Reduction of the violation of animal rights
6. Reduction of food waste

[14] See the appendix section A4.2 to get access to the full coding scheme.

Table 12: Category scheme regarding the understanding of sustainable food consumption[14]

Main Category	Sub-category 1	Sub-category 2	Sub-category 3	Mentions n	Mentions m	Participants
2 Attitude	2.2 Underlying beliefs	2.2.1 Understanding of Sustainable food consumption	2.2.1.1 Reduction of consumption of products with excessive packaging (especially plastic)	9	13	P1.I, P2.I, P3.I, P4.I, P5.I, P6.I, P7.I, P9.I, P10.I
			2.2.1.2 Regional origin of the products	8	13	P1.I, P2.I, P3.I, P4.I, P6.I, P7.I, P9.I, P10.I
			2.2.1.3 Abstinence from meat products	7	7	P1.I, P2.I, P3.I, P4.I, P6.I, P7.I, P9.I, P10.I
			2.2.1.4 Focus on naturally produced/ organic products	8	11	P1.I, P2.I, P4.I, P6.I, P7.I, P8.I, P9.I, P10.I
			2.2.1.5 Low C02 footprint	5	8	P3.I, P6.I, P7.I, P8.I, P10.I
			2.2.1.6 Low food waste	5	6	P1.I, P5.I, P6.I, P7.I, P9.I
			2.2.1.7 Reduction of consumption of products violating animal rights	3	4	P7.I, P8.I, P10.I
			2.2.1.8 Reduction of consumption of products that are not fair traded	3	3	P7.I, P8.I, P10.I
			2.2.1.9 Reduction of the consumption of products with excessive water use	2	2	P1.I, P7.I
			2.2.1.10 Abstinence from dairy products	1	1	P4.I

Own representation

7. Reduction of unhealthy / inorganic products

Expected Consequences of Unsustainable Food Consumption

In order to understand the attitudes of the participants better and why they formed, it is important to capture the consumers' expected consequences of engaging respectively not engaging in sustainable food consumption. Therefore, the participants were asked what consequences they expected if humanity keeps consuming unsustainably.

The participants mostly fear ecological consequences from unsustainable food consumption. The mostl mentioned consequence was a accelerated climate change (P1.I; P3.I; P4.I; P5.I; P6.I; P8.I; P9.I; P10.I). Furthermore, Participant 4 stated that he expects that "the primeval forests are being cut down even faster" (P4.I) and Participant 1 added his concern for the rainforest, which was also supported by the opinion of Participant 9 (P1.I; P9.I).

Another frequently mentioned expected consequence was the rising amount of trash and the environmental pollution (P4.I; P9.I; P10.I). Especially affected are the oceans by getting polluted with plastic according to Participant 4 (P4.I). Besides environmental effects, the participants also predict some humanitarian impacts. Participant 7 talked about possible streams of refugees (P7.I), while Participant 9 sees shortages in resources like food and water coming (P9.I).

Generally, a big part of the participants do not think that they will be directly affected by the variety of consequences (P4.I; P5.I; P7.I; P8.I). Participant 8 e.g. stated: "Personally, I think I'd come off pretty well" (P5.I, own translation), but adds that "the following generations would rather have a bigger problem" (P5.I, own translation), an opinion that participant 4 shares (P4.I). Participant 7 and Participant 8 rather expect other people to suffer from the consequences (P7.I; P8.I). Participant 7 mentioned that indigenous and African people will suffer from the rise in global temperature, especially from heat waves and from water shortages (P7.I). Participant 8 additionally mentions the inhabitant of island countries, whose islands are endangered by the rise in sea-level (P8.I). This could indeed result in refugee streams from these areas as previously mentioned.

Table 13 provides the category scheme regarding the expected consequences, giving an overview of the number of statements assignable to the respective category (n) as well as the total number of mentions (m) and the respective participants.

Table 13: Category scheme regarding expected consequences[15]

Main Category	Sub-category 1	Sub-category 2	Sub-category 3	Sub-category 4	Mentions n	Mentions m	Participants
2 Attitude	2.2 Underlying beliefs	2.2.2 Expected consequences of unsustainable food consumption	2.2.2.1 Ecological consequences	2.2.2.1.1 Facilitating the climate change	8	8	P1.I, P3.I, P4.I, P5.I, P6.I, P8.I, P9.I, P10.I,
				2.2.2.1.2 Deforestation	3	3	P1.I, P4.I, P9.I
				2.2.2.1.3 Detoriation of air quality	1	1	P1.I
				2.2.2.1.4 Trash / pollution	3	4	P4.I, P9.I, P10.I
			2.2.2.2 Humanitarian consequences	2.2.2.2.1 Refugee streams	1	1	P7.I
				2.2.2.2.2 Resource shortage	1	2	P9.I
			2.2.2.3 Concerned people	2.2.2.3.1 No big impact on one-self	4	5	P4.I, P5.I, P7.I, P8.I
				2.2.2.3.2 Descendants will directly feel consequences	2	2	P4.I, P5.I
				2.2.2.3.3 Other peoples will feel the consequences	2	3	P7.I, P8.I

Own representation

In summary, the participants expect some serious consequences. However, the bigger part of them does not think to be heavily affected by these influences themselves. They rather expect the consequences to come at a later point in time, affecting only future generations. Also, they are convinced that it will affect African and island state peoples in particular.

Scepticism and Negative Beliefs

Some of the questioned people also talked about holding rather sceptical thoughts regarding sustainable food consumption. For example, Participants 5 and 8 think that the previously described consequences from unsustainable food consumption cannot be stopped (P5.I; P8.I), which may be related to the belief of some participants that not all people will take part in consuming sustainably (P2.I; P3.I; P5.I). One interviewee even stated that he is not entirely convinced of the current sustainability trend in general. He described himself as sceptical regarding "what is told, written and posted" (P5.I, own translation) by other people and the media (P5.I). Furthermore, he is also convinced that other countries are not as concerned about sustainability, especially when it comes to plastic packaging, and "they are not living in another world" (P5.I, own translation).

The category scheme regarding skeptical and negative beliefs, providing an overview of the number of statements assignable to the respective category (n) as well as the total number of mentions (m) and the respective participants is displayed below in table 14.

4.1.3. Subjective Norm

This chapter examines how the social environment of the participants influences them in their consumer behavior. Most of the interviewees said that they are positively influenced by their friends, respectively acquaintances of their own age, meaning other Millennials (P1.I; P2.I; P3.I; P4.I; P7.I; P8.I; P9.I; P10.I). Participant 4 said that he talks to his friends about necessary changes in consumer behavior (P4.I), while Participant 1 mentioned discussions with fellow students as a major influence (P1.I). Participant 10 explained how many of her friends tell her about their efforts regarding sustainability in general and how this incentivizes her to act accordingly (P10.I). Only one interviewee claimed to be

[15] See the appendix section A4.2 to get access to the full coding scheme.

[16] See the appendix section A4.2 to get access to the full coding scheme.

Table 14: Category scheme regarding skepticism and negative beliefs[16]

Main Category	Sub-category 1	Sub-category 2	Sub-category 3	Mentions n	Mentions m	Participants
2 Attitude	2.2 Underlying beliefs	2.2.3 Scepticism and negative beliefs	2.2.3.1 Negative impacts cannot be stopped	2	2	P5.I, P8.I
			2.2.3.2 Not all people will participate	3	4	P2.I, P3.I, P5.I
			2.2.3.3 Sceptical about the sustainability trend in general	1	2	P5.I
			2.2.3.4 In other countries sustainability has no significance	1	1	P5.I

Own representation

negatively influenced by friends, stating that said she has to let herself "be nudged a little bit negatively" (P6.I, own translation).

The life partner seems to play an important role as four of the questioned persons mentioned that they are positively influenced by their partners (P1.I; P3.I; P6.I; P7.I). Participant 1 explained the fact that he became a vegetarian by the fact that his girlfriend had been a vegetarian before (P1.I). Similarly, Participant 7 is eating less meat and meat products as a result of his girlfriend not eating meat at all (P7.I). Even Participant 6, who claimed to have a rather neutral attitude regarding sustainable food consumption, stated that she is talking to her boyfriend a lot about consumption in general (P6.I). The fact that the life partner exerts a big influence makes a lot of sense, as many of the participants live together with them, so their food consumption is directly dependent on the partners' opinion and consumer behavior.

The third important group of people who have an influence on the participants consumer behavior is the family. However, it is noticeable that the participants rather talked about a negative influence (P1.I; P6.I; P7.I; P8.I; P10.I). Participant 8 states that climate change is not really a topic his parents are concerned with (P8.I). Participant 7 mentioned that members of his family are not really reflecting on their consumer behavior and just buy whatever comes to their mind. He even added that his grandmother and his father would be disappointed if he did not eat meat at family celebrations (P7.I). Participant 6 described her family's attitude regarding sustainability as "weird" (P6.I, own translation). Only one of the interviewees stated that she is positively influenced by her family but explicitly mentioned her sons as the pivotal family members (P2.I). So again, this is a younger generation exerting positive influence.

Furthermore, two participants feel like they are actively influencing other people regarding sustainable consumption (P7.I; P8.I). Participant 7 explained how he influences his family and vice versa, describing it as a "tug-of-war, that is working in both directions" (P7.I, own translation). Only one participant stated to consume independently and not being influenced by his social environment at all (P5.I).

Table 15 provides the category scheme regarding the subjective norm, giving an overview of the number of statements assignable to the respective category (n) as well as the total number of mentions (m) and the respective participants.

Summary

All in all, it seems like the generation of the Millennials themselves exert a strong positive influence on each other, while older generations seem to have a rather negative influence on them, which, on the other hand, is often rather passive. Many participants stated that the positive influence affects them more than the negative ones (P6; P10). In sum, the influence of the social environment seems to be positive.

4.1.4. Perceived Behavioral Control

Participants feel like they do not have full control over their behavior due to a variety of reasons. The most frequently mentioned reason is the higher price of sustainable food products (P1.I; P2.I; P3.I; P4.I; P5.I; P6.I; P7.I; P8.I; P9.I). Participants 1 and 9 for example explained that since they are still students, they cannot afford to buy their daily food products in an organic supermarket and rather need to keep their expenses as low as possible (P1.I; P9.I). Participant 7 stated that the price level in an organic supermarket is unproportionally high, implying that he would pay more for more sustainable products but not as much as the organic supermarkets charge at the moment (P7.I). Sometimes food prices in these stores are many times higher than those in regular supermarkets according to Participant 8 (P8.I).

Another frequently mentioned issue is the fact, that many of the participants feel that they are lacking the knowledge to really assess the sustainability of the various products (P1.I; P3.I; P4.I; P6.I; P7.I; P8.I; P9.I). Decisions are made based on gut feeling and estimation according to Participant 8 (P8.I).

[17] See the appendix section A4.2 to get access to the full coding scheme.

Table 15: Category scheme regarding the subjective norm[17]

Main Category	Sub-category 1	Sub-category 2	Mentions n	Mentions m	Participants
3 Subjective Norm	3.1 Friends/ Millennials	3.1.1 Positive influence	8	13	P1.I, P2.I, P3.I, P4.I, P6.I, P8.I, P9.I, P10.I
		3.1.2 Negative influence	1	1	P6.I
	3.2 Life partner	3.2.1 Positive influence	4	4	P1.I, P3.I, P6.I, P7.I
		3.2.2 Negative influence	2	2	P2.I, P10.I
	3.3 Family	3.3.1 Positive influence	1	1	P2.I
		3.3.2 Negative influence	5	6	P1.I, P6.I, P7.I, P8.I, P10.I
	3.4 Active exertion of influence on others		2	2	P7.I, P8.I
	3.5 No influence on social environment		1	1	P5.I

Own representation

Participant 1 explained that the topic of sustainability regarding food consumption, in general, is hard to "penetrate" (P1.I, own translation) and Participant 9 claimed that he does not really know how to distinguish between sustainable and unsustainable products (P9.I). Participant 6 blames the school system and demands a school subject that deals with sustainable food consumption (P6.I). Participant 7 even added that he would like "someone else to do the thinking" for him (P7.I, own translation), expressing that he does not even want to deal with the topic in more detail.

Also, many participants complained about the lack of sustainable products in regular supermarkets (P1.I; P2.I; P3.I; P5.I; P6.I; P7.I; P10.I), especially regarding plastic-free packaged products (P1.I; P2.I; P3.I). He tries to buy plastic-free, but it is simply not possible said one interviewee (P3.I). Participant 2 stated that sometimes there are simply no regionally produced alternatives and mentioned sugar snaps as an example, which she needed for a certain recipe (P2.I). When confronted with the fact that a possible solution would be to change the store they are shopping in, many of the interviewees stated reasons why that is not an option. One of them, who is living in a smaller city, explained that shopping facilities are very limited (P6.I). Another main reason appears to be the fact, that there is no store that offers all products the participants need (P2.I; P4.I; P6.I). Participant 6 complained that she would have to visit ten different stores in order to get all food products she needs (P6.I). This corresponds to another frequently mentioned aspect, which is a lack of time (P1.I; P2.I; P3.I; P5.I; P6.I). The participants do not have unlimited time to get their grocery shopping done, so they cannot afford to visit many different stores.

Another major reason that a variety of interviewees mentioned, was a lack of trust in the variety of different sustainability labels, on the one hand, and the supermarket chains (P1.I; P2.I; P3.I; P5.I; P6.I; P7.I; P10.I) and food manufacturers (P5.I) on the other hand. Participant 4 as an example describes fair trade as an "obscure thing" (P4.I, own translation) that he does not really trust. "I sometimes find it difficult to really judge whether a product that is labeled sustainable or ecological has really been produced sustainably" said Participant 9. Another interviewee went even further by explaining that food manufacturers often "work with slogans on packaging to pretend sustainability" (P5.I, own translation). He suspects a lot of hypocrisy and eye washing in that context (P5.I). Especially for the big food corporations and supermarket chains, he thinks that it is all about the money and not about sustainability (P5.I). At that point, it is interesting to mention, that the overall education did not seem to play any role regarding the knowledge about sustainable food products. The interviewees who mentioned this lack of trust, have different educational backgrounds ranging from vocational training to an academic Master degree (see section 3.1 for details).

The category scheme regarding the PBC, providing an overview of the number of statements assignable to the respective category (n) as well as the total number of mentions (m) and the respective participants, is displayed below in table 16.

Summary

Overall, there is a variety of reasons that leads to the participants' perception that they are not in full control of their behavior regarding sustainable food consumption. Some of the reasons, however, are based on the same main issues. The lack of trust in the sustainability labels for example is a

[18] See the appendix section A4.2 to get access to the full coding scheme.

Table 16: Category scheme regarding Perceived Behavioral Control[18]

Main Category	Sub-category 1	Sub-category 2	Mentions n	Mentions m	Participants
4 Perceived Behavioral Control	4.1 Price		9	17	P1.I, P2.I, P3.I, P4.I, P5.I, P6.I, P7.I, P8.I, P9.I
	4.2 Lack of knowledge/ education		7	15	P1.I, P3.I, P6.I, P7.I, P8.I, P9.I, P10.I
	4.3 Lack of sustainable food offers	4.3.1 General	7	12	P1.I, P2.I, P3.I, P5.I, P6.I, P7.I, P10.I
		4.3.2 Products with plastic-free packaging	3	3	P1.I, P2.I, P3.I
	4.4 Lack of trust	4.4.1 In sustainability label	5	7	P1.I, P2.I, P3.I, P5.I, P6.I, P7.I, P10.I
		4.4.2 In supermarkets/ big food manufacturers	1	3	P5.I
	4.5 Lack of time		5	7	P1.I, P2.I, P3.I, P5.I, P6.I
	4.6 No shop that offers all the sustainable products needed		3	4	P2.I, P4.I, P6.I

Own representation

result of the fact that the participants did not really deal with these labels and can therefore not distinguish the real sustainability labels from the greenwashing labels. This aspect can be subsumed under "lack of knowledge / education". If people are familiar with the labels and know which of them are trustworthy, it is not necessary to trust the food corporations or supermarket chains as people could simply rely on the respective labels. The fact that there is no store supplying all sustainable products, is only a problem as the participants lack the time to visit multiple stores, which is the actual obstacle here. This leads to the following main external obstacles preventing people from consuming sustainably:

1. Price
2. Lack of knowledge/education
3. Lack of sustainable product offers
4. Lack of time

4.1.5. Self-efficacy

A big part of the questioned people blamed their own laziness or convenience for not actually consuming as sustainable as they actually could (P1.I; P2.I; P3.I; P4.I; P6.I; P7.I; P8.I). Participant 4 stated that due to his "weaker self" (P4, own translation") he wants to buy everything in one store, leading him to go to a regular food store rather than to an organic supermarket (P4.I). Participant 7 added that the organic supermarket close to his home has inconvenient opening hours (P7.I). Because of his laziness, he is not dealing with sustainable food consumption and has therefore not enough knowledge to identify sustainable products (as described in sub-chapter 4.1.4), said one of the interviewees (P8.I).

Also frequently mentioned by the participants was that the purchase decisions were rather unconscious, respectively decisions born out of habit (P1; P3; P4; P5; P8; P9; P10.I). Participant 8 mentioned a "basic assortment" (P8.I, own translation), which he buys again and again, because he knows these products well and where to get them in the store, so he can "deal with this subject quickly" (P8.I, own translation). He added that he keeps on buying one specific cheese, even though he does not even like it, out of pure habit. In this context, Participant 3 rather talked about how changing one's habit is hard as she thinks that "humans are creatures of habit" (P3.I, own translation) and that it takes time. She was referring to the fact that she started to change her consumer behavior some years ago as previously described in sub-chapter 4.1.2.1.

Some of the interviewees implied, that sometimes they buy unsustainable products simply because of a desire for these products (P1.I; P2.I; P4.I; P5.I; P9.I; P10.I). Participant 2 for example said, that she sometimes just "gets weak" (P2.I, own translation), while Participant 5 mentioned that his favorite candy is very unsustainable, but he just likes it so much (P5.I).

Another aspect is closely connected to the previously mentioned lack of time. Several participants stated that sometimes the reason they buy unsustainable products is

their own impatience (P1.I; P3.I; P4.I). Firstly, this refers to the fact that they only want to go to one store to get everything they need (P4.I), as mentioned in chapter 4.1.4. Secondly, it also refers to participants not taking much time for their shopping trip and therefore choosing products without really dealing with them (P1.I; P3.I). In that context, Participant 1 added he wants to minimize the time spent in the store as he feels uncomfortable while shopping (P1.I).

Table 17 provides the category scheme regarding self-efficacy, giving an overview of the number of statements assignable to the respective category (n) as well as the total number of mentions (m) and the respective participants.

Summary

Overall, there is a variety of internal obstacles preventing people from consuming according to their attitude. Again, some obstacles may be a result of the same basic issues. Correspondingly, participants' laziness and convenience are connected to their impatience and the fact that they may feel uncomfortable while shopping. All these factors lead people to minimizing the time spend on grocery shopping. This leads to the following three main internal obstacles:

1. Minimizing the time spent for grocery shopping
2. Unconscious decision/habit
3. Desire for unsustainable products

4.2. Observation and the Verbal Protocol Analysis

The first chapter of this section presents the insights of the observation as captured by the researcher. After that, the second chapter is dealing with the thoughts of the participants captured by the VPA. The sources for the statements in that section are the observation combined transcripts from the respective participants. For example, the observation combined transcript of Participant 1 is shortened as "P1.O".

4.2.1. Observation

During the observation, it was most striking that most of the purchases (m=57) did not include an observable decision process. It seemed like the participants just took the respective products without even really looking at them or considering alternative products, probably because they were already familiar with the product. Often it appeared as if the actual purchase decision had been made earlier and in the store, the purchase is only carried out. It is important to mention that this could be observed multiple times during the shopping trips of all ten participants (P1.O; P2.O; P3.O; P4.O; P5.O; P6.O; P7.O; P8.O; P9.O; P10.O).

Sometimes the observed persons did not even stop to pick a product but rather just put them into the shopping cart or bag while passing by (m=22). Often the purchases seemed to be spontaneous. A major part of the participants showed this behavior (P2.O; P4.O; P6.O; P7.O; P8.O; P9.O; P10.O).

Often the participants took a superficial look at the product before picking it (m=33). Participants seemed to evaluate its appearance, e.g. regarding its freshness or different flavors within the same brand but did not properly read the information on the packaging respectively really thinking about the product. Here as well, it is important to mention that this behavior could be observed multiple times during the shopping trips of all participants (P1.O; P2.O; P3.O; P4.O; P5.O; P6.O; P7.O; P8.O; P9.O; P10.O).

Only six participants (P1.O; P2.O; P3.O; P7.O; P9.O; P10.O) appeared to even consider alternatives at least a few times during their shopping trips (m=13). However, in many of these cases, participants seemed to not really think about the products in detail, but again rather compare superficial attributes like appearance (P2.O; P3.O) or the quantity in the package (P2.O). Only selectively the participants seem to actively think about the products and evaluate alternatives (P3.O; P7.O; P10.O).

Furthermore, it was striking that many participants seemed to reflect or even hesitate where to go next, respectively were actively searching for a product at some point in the shopping trip (P1.O; P3.O; P4.O; P5.O; P9.O; P10.O). In that context, it is to be noted that especially the two participants that claimed to go to the same store every time, so they are familiar with it and know how to find each product, showed this insecurity (P1.O; P5.O).

The category scheme regarding the observed behavior, providing an overview of the number of statements assignable to the respective category (n) as well as the total number of mentions (m) and the respective participants, is displayed below in table 18.

Summary

All in all, it can be stated that regarding the major part of the products, it seemed that the actual purchase decision was made at some point in the past. The observed behavior during the shopping trip appeared to be mostly defined by previously formed habits.

4.2.2. Verbal Protocol Analysis

Reasons for Product Choices

By far the biggest share of the participants' thoughts obviously centered around the products they intend to buy and why they do so. It is striking that in most cases participants (P1.O; P2.O; P3.O; P4.O; P5.O; P6.O; P7.O; P8.O; P9.O; P10.O) only stated that they need a certain product, without thinking about why they need it, what alternative products could there be or any other additional considerations (m=82). These thoughts are all similar like: "now I am looking for cheese" (P1.O, own translation), "I need butter" (P10.O, own translation) or "Now I am going to get some grapes" (P2.O, own translation). This corresponds well with the impression from the observation that the actual purchase decision had taken place before the shopping trip, which then only marked the realization of the purchases.

[19] See the appendix section A4.2 to get access to the full coding scheme.

[20] See the appendix section A4.2 to get access to the full coding scheme.

Table 17: Category scheme regarding self-efficacy[19]

Main Category	Sub-category 1	Mentions n	Mentions m	Participants
5 Self-efficacy	5.1 Laziness / Convenience	7	19	P1.I, P2.I, P3.I, P4.I, P6.I, P7.I, P8.I
	5.2 Unconscious decisions / habit	7	12	P1.I, P3.I, P4.I, P7.I, P8.I, P9.I, P10.I
	5.3 Desire for unsustainable products	6	6	P1.I, P2.I, P4.I, P5.I, P9.I, P10.I
	5.4 Impatience / Taking too little time	3	4	P1.I, P3.I, P4.I
	5.5 Uncomfortable shopping	1	1	P1.I

Own representation

Table 18: Category scheme regarding the observation[20]

Main Category	Sub-category 1	Sub-category 2	Mentions n	Mentions m	Participants
6 Behavior during shopping	6.1 Observation	6.1.1 Selection without closer examination of the products	10	57	P1.O, P2.O, P3.O, P4.O, P5.O, P6.O, P7.O, P8.O, P9.O, P10.O
		6.1.2 Superficial view at the products	10	33	P1.O, P2.O, P3.O, P4.O, P5.O, P6.O, P7.O, P8.O, P9.O, P10.O
		6.1.3 Taking the product in passing	7	22	P2.O, P4.O, P6.O, P7.O, P8.O, P9.O, P10.O
		6.1.4 Dealing with the product and alternatives	6	13	P1.O, P2.O, P3.O, P7.O, P9.O, P10.O
		6.1.5 Reflecting/hesitating	5	8	P1.O, P3.O, P4.O, P5.O, P9.O, P10.O

Own representation

Also, almost all participants frequently thought about what they like or dislike regarding taste respectively about a general desire for certain products (P1.O; P2.O; P3.O; P4.O; P5.O; P6.O; P7.O; P9.O; P10.O). For example, Participant 1 chose a pack of tortellini after having stated “I’m in the mood for tortellini again”. (P1.O, own translation) or Participant 2 who did not get a certain salad as she is “has grown tired of this one” (P2.O, own translation).

Another relevant category of thoughts was about the lacking offers in the respective stores or lacking alternative products (P1.O; P2.O; P3.O; P4.O; P7.O; m=15). This was the case if e.g. a certain product was sold out (P1.O), there was no alternative product without packaging (P2.O; P3.O; P4.O) or if a certain product is not available at all (P7.O).

Often the product choices were directly influenced by the participants’ general social environment (P1.O; P3.O; P7.O; P8.O; m=9). Either because someone actively recommended a certain product (P7.O) or because the participant chooses a product for a person from his social environment, mostly as the partner or flatmate likes a certain product (P1.O; P3.O; P8.O). A lot of the product choices of Participant 10, who is a mother of two, were heavily influenced by the preferences of her children (P10.O).

The impression from the observation, that the participants often examine the appearance of the products could be confirmed by the VPA (P1.O; P2.O; P3.O; P8.O; P9.O; m=13). Frequently, participants stated something like “they look good” (P1, own translation) or “this one does not look good” (P2.O, own translation) when choosing products.

A further driver seems to be the price as implied by the majority of participants (P1.O; P2.O; P4.O; P5.O; P6.O; P7.O; P10.O) in a variety of situations (m=13). For example, Participant 1 said he chooses a certain milk as it is the cheapest (P1.O) or Participant 2 who decided to buy a pack of coconut flakes which was actually too big, but very cheap (P2.O).

In fact, also sustainability plays a role in product choices, however, compared to the previously described factors, it plays a rather minor role (m=7). Correspondingly, Participant 7 tried to buy vegetarian food (P7.O). It is important to mention at that point, that this does not necessarily mean

that her vegetarian consumption is motivated by sustainability. However, it was classified as abstinence from meat like previously mentioned in sub-chapter 4.1.2.2. Participant 4 chose bananas because they were fair traded (P4.O) and Participant 5 chose pasta that was regional (P5.O). Participants 3 and 10 chose certain products because of their sustainable packaging (P3.O; P10.O).

Further thoughts often centered about varying nutrition (P1.O; P5.O; P7.O; P10.O; m =6). Often these lines of thoughts started with "What could we cook next time?" (P1.O, own translation). Participant 10 stated that she chose one product as she never had it before (P10.O).

Special offers influence product choices as well (P2.O; P3.O; P4.O; P6.O; P8.O; P10.O m=6). For example, Participant 4 bought a feta cheese because it was discounted (P4.O) or Participant 3 who stated to generally watch out for special offers (P3.O).

The quantity in a package also seems to be an important aspect as many participants stated due to the VPA (P1.O; P2.O; P4.O; P10.O; m=5). Participant 2, for example, said that she did not choose the fair-trade bananas as they were only available in too big packages (P2.O).

Other only selectively stated aspects included how healthy certain products were (P5.O, P8.O; m=2), some products' best before date (P9.O; m=3), how easy and fast to prepare certain products are (P5.O; m=2) or that the choice was random (P9.O; m=1)

General Thoughts

In accordance with the observation that the participants appeared to be insecure sometimes, they frequently thought about where to find a certain product (P9.O; P10.O) or were reflecting on what they already put in their shopping cart respectively what they already have at home (P2.O; P4.O; P7.O; P8.O; P10.O). Some also considered to buy certain products which they needed somewhere else, either due to an insufficient offer at the current store (P2.O; P4.O) or personal preferences (P4.O; P10.O).

Participant 5 at some point additionally stated that he does not have "the nerve to look for where anything is" (P5.O, own translation), stressing again his dislike for grocery shopping as described in chapter 4.1.1.

Conversation with the Own Child

Participant 6 was a special case as she brought her two-year-old daughter along, for the shopping trip, so she had to deal with her a lot, leaving only little room for her own thoughts. Very often, she asked her daughter whether she liked a certain product and whether she should purchase it (P6.O; m=10) or explained some products to her daughter (P6.O; m=1). Also, some time Participant 6 just talked nonsense with her daughter in order to keep her occupied and in good spirits (P6.O; m=3). Nevertheless, at the end of the shopping trip, her daughter got fretful anyways and she had to finish her shopping early (P6.O).

Table 19 provides the category scheme regarding the VPA, giving an overview of the number of statements assignable to the respective category (n) as well as the total number of mentions (m) and the respective participants.

4.3. Actual Purchase Decisions

It was possible to examine the actual purchase behavior for all participants except for Participant 6. Table 20 presents an overview of the evaluation of purchase decisions per participant and crosschecks it against their self-reported assessment of their attitude as described in sub-chapter 4.1.2.1. The average score regarding the impact from production was calculated as explained in section 3.3. The shares displayed in the table were calculated on the basis of the list of purchased products per participant[22]. An additional letter indicates for each cell if the respective participant mentioned the respective aspect to be an important part of sustainable food consumption, as described in sub-chapter 4.1.2.2. In that context "Y" means that the participant did mention the aspect and "N" indicates that she did not.

In most cases the average score for the impact from production was between 1.5 and 2.0 which means in average the products had a low to medium GHG impact as defined in chapter 2.1.4. This corresponds to the findings from the interviews e.g. that participants tried to avoid meat respectively dairy products (P1.I; P2.I; P3.I; P4.I; P7.I; P9.I; P10.I), which have a particularly high impact (see chapter 2.1.4).

Even though all interviewees except for Participant 5 and 6 said that the origin of the product is an important aspect of sustainable food consumption for them (P1.I; P2.I; P3.I; P4.I; P6.I; P7.I; P9.I; P10.I) and they state to have a positive attitude (P1.I; P2.I; P3.I; P4.I; P7.I; P8.I; P9.I; P10.I) they mostly did not buy regional products. Participants 4 and 5 did at least buy some regional products (P4.O; P5.O), which is remarkable, as the Participant did not even state to consider the origin of the product. Only Participant 10 who went to an organic supermarket bought a substantial part of regional products (P10.O).

The share of unpacked products was generally higher than the share of regional products. Again, Participant 10 had the biggest share, with almost half of the products bought being unpacked. Most of the other participants bought between approximately eight and twenty percent unpacked products (P1.O; P2.O; P3.O; P4.O; P6.O; P7.O; P8.O). Indeed Participant 8, who had the lowest share, did not mention the reduction of packaging as important aspect of sustainable food consumption (P8.O). On the other hand, Participant 5 and Participant 9 both mentioned it but did not even buy one single unpacked product (P5.O; P9.O) and one of them did at least buy one fair-traded product (P10.O).

Only three participants stated fair trade to be an important part of sustainable food consumption (P7.O; P8.O; P10.O). It is remarkable that two of them did not even buy one single fair-traded product (P7.O; P8.O).

Only two participants did not mention natural respectively organic production to be sustainable (P3.O; P5.O). Participant 5 did shop accordingly and bought not even one or-

[21] See the appendix section A4.2 to get access to the full coding scheme.

[22] See the appendix section A3 for a detailed list of purchases per participant

[23] As described in the summary of sub-chapter 4.2.2.1

Table 19: Category scheme regarding the Verbal Protocol Analysis[21]

Main Category	Sub-category 1	Sub-category 2	Sub-category 3	Sub-category 4	Mentions		Participants
					n	m	
6 Behavior during shopping	6.2 Thoughts/ VPA	6.2.1 Reason for purchase	6.2.1.1 "I still need X"/ Habit/ Already determined before		10	82	P1.O, P2.O, P3.O, P4.O, P5.O, P6.O, P7.O, P8.O, P9.O, P10.O
			6.2.1.2 Like / dislike taste		9	27	P1.O, P2.O, P3.O, P4.O, P5.O, P6.O, P7.O, P9.O, P10.O,
			6.2.1.3 Lack of / insufficient offers		5	15	P1.O, P2.O, P3.O, P4.O, P7.O
			6.2.1.4 Influence from social environment	6.2.1.4.1 General	4	9	P1.O, P3.O, P7.O, P8.O
				6.2.1.4.2 Children	1	4	P10.O
			6.2.1.5 Appearance of the products		5	13	P1.O, P2.O, P3.O, P8.O, P9.O
			6.2.1.6 Price		7	12	P1.O, P2.O, P4.O, P5.O, P6.O, P7.O, P10.O
			6.2.1.7 Sustainability	6.2.1.7.1 Plastic / packaging free	2	3	P3.O, P10.O
				6.2.1.7.2 Vegetarian	1	2	P.O
				6.2.1.7.3 Fairtrade	1	1	P4.O
				6.2.1.7.4 Origin of the product	1	1	P5.O
			6.2.1.8 Varied diet		4	6	P1.O, P5.O, P7.O, P10.O
			6.2.1.9 Special offer		6	6	P2.O, P3.O, P4.O, P6.O, P8.O, P10.O
			6.2.1.10 Quantity in the package		4	5	P1.O, P2.O, P4.O, P10.O
			6.2.1.11 Best before date		1	3	P9.O
			6.2.1.12 Healthy		2	2	P5.O, P8.O
			6.2.1.13 Easy to prepare		1	2	P5.O
			6.2.1.14 Random choice		1	1	P9.O
		6.2.2 General thoughts	6.2.2.1 "I already have product X"		5	14	P2.O, P4.O, P7.O, P8.O, P10.O
			6.2.2.2 "Where is product X"		2	6	P9.O, P10.O
			6.2.2.3 "I will buy product x somewhere else"		3	5	P2.O, P4.O, P10.O
			6.2.2.4 No desire to shop		1	1	P5.O
		6.2.3 Conversation with the own child	6.2.3.1 What the child wants		1	10	P6.O
			6.2.3.2 Explanation of products		1	1	P6.O
			6.2.3.3 Unrelated		1	3	P6.O

Own representation

Table 20: Crosscheck of participants' self-report and the actual purchased products

Participant	Overall attitude	Average score impact from production	Share of regional products	Share of unpacked products	Share of fair-traded products	Share of organic products
		Did the participant mention the respective aspect or related ones[23] as an important part of sustainable food consumption? Y = Yes N = No				
1	Unlimited positive	1.69 Y	0.00% Y	10.00% Y	0.00% N	0.00% Y
2	Unlimited positive	1.53 Y	0.00% Y	15.60% Y	0.00% N	0.00% Y
3	Unlimited positive	1.38 Y	0.00% Y	20.00% Y	0.00% N	16.67% N
4	Limited positive	1.83 Y	4.00% Y	12.00% Y	4.00% N	4.00% Y
5	Limited positive	1.81 N	7.14% N	0.00% Y	0.00% N	0.00% N
7	Limited positive	1.9 Y	0.00% Y	14.30% Y	0.00% Y	14.30% Y
8	Limited positive	1.7 Y	0.00% N	8.57% N	0.00% Y	5.71% Y
9	Limited positive	1.73 Y	0.00% Y	0.00% Y	0.00% N	0.00% Y
10	Unlimited positive	1.55 Y	18.18% Y	45.45% Y	4.54% Y	100% Y

Own representation

ganic product. Participant 3 in contrast, bought more than 16% organic products, only second to Participant 10, who bought organic products only, as she visited an organic supermarket. The other participant did not buy or did only buy very few organic products even though they claimed it to be important for sustainable food consumption (P1.O; P2.O; P4.O; P7.O; P8.O; P9.O).

In summary, the actual purchase decisions of the participants did not really reflect the self-report given during the interviews. The only exception is Participant 10 who did in fact buy many organic, regional, and unpacked products (P10.O).

5. Summary and Limitations

5.1. Summary

This thesis picked up on the recent trend of sustainability and sustainable consumption by investigating the sustainable food consumption of German Millennials regarding an assumed gap between their attitude and their actual consumer behavior. Based on the TPB, which served as the theoretical framework of the thesis, a multi-step, qualitative research process was developed in order to find viable explanations for the ABG. The process included the triangulation of questioning everyday consumers via qualitative depth interviews, the observation of those consumers during one of their regular grocery shopping trips, and the superficial analysis of the products they bought. The summary of the study's main findings is described in the following two chapters.

5.1.1. Occurrence of the Attitude Behavior Gap

There seems in fact to be a gap between the sustainable attitude and the actual behavior. On the one hand, half of the participants even claim to still have room for improvement as described in sub-chapter 4.1.2.1. On the other hand, most participants have a clear positive attitude regarding sustainable food consumption, which was previously described in sub-chapter 4.1.2.1 as well. However, sections 4.2 and 4.3 showed as well that the actual behavior of the participants did in fact not correspond completely with that positive attitude. Many participants did often buy products with higher impacts regarding the evaluation scheme developed in chapter 2.1.4 and only selectively bought fair traded products (see section 4.3). Additionally, the results of the VPA suggest that sustainability only plays a minor role during the actual shopping trip (see chapter 4.2.2). The results presented in section 4.3 imply a gap between the participants' self-reports and their actual behavior as well.

5.1.2. Explanations of the Attitude Behavior Gap

In sum, the influence of the social environment was positive as chapter 4.1.3 showed and should therefore not provide an explanation for the ABG. However, participants men-

tioned several factors regarding PBC and self-efficacy detaining them from fully consuming sustainably and thus providing possible explanations for the ABG. Thus, these determinants are examined in the following sub-chapters.

Perceived Behavioral Control

High Prices of Sustainable Products

The most frequently mentioned external obstacle was the high price level of sustainable food products, implying that people may not be able to afford these products (P1.I; P2.I; P3.I; P4.I; P5.I; P6.I; P7.I; P8.I; P9.I). Especially for people who may have a limited budget like e.g. students, this could be a valid explanation. However, only one participant stated to have a limited budget for food (P1.O). All others participants said they can theoretically spend as much as they like for food (P2.O; P3.O; P4.O; P5.O; P6.O; P7.O; P8.O; P9.O; P10.O). This indicates that even though participants are actually able to afford sustainable food, they just lack in willingness to spend the higher prices compared to the other products.

Lack of Knowledge / Education Regarding Sustainable Products

In some cases, the ABG can be a result of a lack of knowledge. That means people may want to consume according to their attitude but fail to do so, as they are unable to identify sustainable respectively unsustainable products and therefore simply fail to choose sustainable products (P1.I; P3.I; P4.I; P6.I; P7.I; P8.I; P9.I).

Insufficient Sustainable Food Product Range

Another possible obstacle is the lacking range of sustainable products. Especially in the regular supermarkets like Aldi, Lidl or Rewe, there is only a poor offer of unpacked product. On the other hand, people could also choose to shop at organic supermarkets or the weekly farmers' market that definitely provide a sufficient offer of sustainable products. However, the participants stated various reasons why they do not choose to shop there, above all, the higher prices and the expectation to not getting everything they need at one store (P1.I; P2.I; P3.I; P5.I; P6.I; P7.I; P10.I).

Self-efficacy

Minimizing the Time Spent for Grocery Shopping

One important internal obstacle preventing participants from consuming according to their attitude is peoples' laziness and their tendency to be convenient (P1.I; P2.I; P3.I; P4.I; P6.I; P7.I; P8.I). That means people are too lazy to deal with the subject sustainability in order to fill their knowledge gaps. Also, people like the convenience of getting all products needed in one store, which is why they keep shopping in regular supermarkets that might have an adequate offer of sustainable products (as described in the previous sub-chapter). Additionally, many participants claimed to be impatient (P1.I; P3.I; P4.I). All in all, that means people are trying to minimize the time and effort they put into grocery shopping, possibly leading to hasty and spontaneous purchases, which corresponded with the findings from the observation and the VPA (see section 4.2). As they are often in a hurry, they do not really think about why they choose a certain product but rather do it automatically as confirmed by the observation and VPA (P1.O; P2.O; P3.O; P4.O; P5.O; P6.O; P7.O; P8.O; P9.O; P10.O). This may result in unthoughtful and unconscious product selection that is not reflecting the consumer's attitude.

Unconscious Decisions Respectively Habit

Also, participants frequently stated that many of their purchase decisions are unconscious or habitual (P1.I; P3.I; P4.I; P5.I; P8.I; P9.I; P10.I) which corresponds to the findings from the observation and the VPA (see section 4.2). This means the actual decision was made before the shopping trip, which is only the realization of that decision. In combination with the fact that some participants said that they changed their attitude in the last years, a possible explanation for the ABG is that people simply di not have enough time to change their habits.

Desire for Unsustainable Products

Many participants mentioned that they sometimes buy unsustainable products out of desire. However, this is only the case for single exceptions (P1.I; P2.I; P4.I; P5.I; P9.I; P10.I). Accordingly, this aspect does probably not provide a suitable explanation for the ABG, as the gap does not only refer to single purchase behaviors but rather the overall consumer behavior.

5.2. Limitations

General limitations regarding all qualitative research, e.g. regarding intersubjectivity or generalizability have to be kept in mind in any case (Flick, 2018; Giri & Biswas, 2019). Furthermore, there are specific limitations of this thesis, which will be discussed in this section.

5.2.1. Theoretical Framework

On the one hand, one of the main findings of the study is that sustainable food consumption is strongly determined by habit. On the other hand, the TPB was previously often discussed regarding its validity in the context of habitual behavior (see sub-chapter 2.2.2.3). Possibly, another framework, for example, one that explicitly conceptualizes the habitual aspect of behavior would have been more suitable for this thesis.

The SHIFT model

This model has recently been used to encourage a change of behavior to act more sustainably (White, Habib, & Dahl, 2020; White, Habib, & Hardisty, 2019). The authors explicitly state that the framework can help to address the attitude behavior gap so it could have also been a suitable approach to explain the attitude behavior gap. It includes five key drivers of behavior change regarding sustainability: social influence, habit formation, the individual self, feelings and cognition, and tangibility (White et al., 2020).

The "Social Influence" is similar to the subjective norm from the TPB, involving beliefs about what is socially appropriate or expected.

The "Individual Self" depicts a group of factors centered on the actual individual, who desire to have a positive self-view (White et al., 2019). This can lead to positive outcomes (Dunning, 2007) or to negative outcomes as a result of self-defensive reactions (Dickinson, 2009). Self-consistency describes that in addition to the fact that people want to see themselves positively, they also want to see themselves as ben consistent. Self-consistency research suggests that consumers engaging in sustainable behavior at one point in time often leads to consistent sustainable behavior in the future (White et al., 2019). Also, self-interest plays a major role in influencing behavior according to White et al. (2019) as well as self-efficacy which they base on the work of (Bandura & Walters, 1977). Also, individual differences that are similar to the "external factors" as described in subchapter 2.2.1.1, play a role according to the SHIFT model (White et al., 2019). In summary, it can be stated, that many of the concepts subsumed under the "individual self" can also be found within the extended TPB as described in this chapter.

The key factor "Feelings and Cognition" describes on the one hand how positive (e.g. love for nature or pride) and negative emotions (e.g. guilt or fear) can heavily influence consumer decisions and are therefore a major part of the SHIFT model. On the other hand, it also includes how information, learning, and knowledge as important aspects (White et al., 2019). This coincides with one of the major findings from this thesis, the lack of knowledge as a driver influencing PBC and therefore a reason for the ABG.

White et al. (2019) also introduce some rather new aspects like the key factor "Tangibility". This depicts the unique facet of sustainable consumption that respective actions and outcomes often seem abstract, vague, and distant from oneself. If respective outcomes are communicated clearer, more concrete, and more comprehensive, people are more likely to consume sustainably.

Most importantly, "Habit Formation" is stated as one key driver (White et al., 2019). That includes the breaking of bad habits as well as the actual formation of desirable habits. Retrospective, as habits emerged as a main driver for the ABG in the context of sustainable food consumption of German Millennials the SHIFT model may have been a more suitable guiding framework in order to explore the gap than the TPB.

On the other hand, the framework originally refers to changing consumer behavior and not explaining it (White et al., 2019), which was not the scope of this thesis. That is why the TPB was chosen over the SHIFT model for this thesis.

Value-belief-norm-theory

Another possible theory that is frequently mentioned in the context of pro-sustainable behavior is the Value-belief-norm-theory. It does not conceptualize habits as well, but rather focuses on a person's values (Stern, 2000). It did not have obvious advantages over the TPB with regard to explaining and exploring sustainable consumer behavior. However, it was less suitable to illustrate the ABG, which is why it was not chosen as guiding framework of this thesis.

Conclusion

In Summary, the TPB illustrates the ABG well and was designed to explore and explain behavior. It was a sound theoretical framework for this thesis and succeeded in providing several sensible explanation approaches for the ABG. However, for further research in the domain of sustainable consumption, a further extension of the TPB is needed in order to better capture habits respectively the processual and repeating character of sustainable food consumption.

5.2.2. Research Design

Sample

The sample is especially composed of people at least indirectly personally known to the researcher. As the sample was not randomly chosen, this might lead to a selection bias and therefore possibly limiting the generalizability of results (Atteslander, 2013; Geddes, 1990).

Furthermore, there was only a limited geographical spread, as only participants from the south of Germany were selected. Including participants from other parts of Germany may have added some additional insights. However, it is questionable if people from other parts of Germany really differ considerably in their consumer behavior.

Content Analysis

The content analysis followed the approach by Mayring (2010) but the involvement of just one researcher in the process did not offer the possibility to assess the inter-coder reliability.

5.2.3. Results

Technical Fails During Shopping Trips

Like described in chapter 3.2.3, the audio gear failed in three cases (P6.O; P7.O; P8.O). A considerable part of the information, that could have been captured with the audio record was lost for these participants. In these cases, a protocol was created from the fresh memories of the researcher, still ensuring insights into these participants' shopping behavior and partially compensate for this loss.

Possible Bias

Additionally, results may be distorted e.g. due to the social desirability bias (Atteslander, 2013; Nederhof, 1985), which reflects the tendency on behalf of the participants to deny socially undesirable traits and behaviors and to claim socially desirable ones, and the tendency to say certain things in order to appear in a favorable light (Atteslander, 2013; Nederhof, 1985). In order to lower the bias, the study also included an observation part in addition to the self-report in the interviews.

Evaluation of the Purchase Decisions

The evaluation of the purchase decisions regarding sustainability is done on a high level and was not developed within a scientifically sound process. Therefore, the evaluation provides only a limited value. However, the quantification was explicitly not the main objective of the thesis. The evaluation was only used to roughly crosscheck the participants' self-report and their actual purchase decisions and in order to be able to compare the purchase decisions of the different participants.

6. Implications, Contribution, and Further Research

The last part describes the practical implications, which can be derived from the key findings of the thesis provides as well as the theoretical contribution and possible directions for further research.

6.1. Practical Implications

6.1.1. Sustainable Food Manufacturers and Supermarket Chains

This thesis provides insights into the decision-making of consumers and contributes to a better understanding of the ABG gap within the food industry. The presented findings might be of potential interest to sustainable food manufacturers and supermarket chains, who are keen to transform consumers' sustainable attitudes into actual buying behavior. Understanding the consumers' attitude towards sustainable consumption as well as their constraints regarding PBC and self-efficacy may help the industry to the following issues.

Participants complained about the insufficient range of sustainable products (see chapter 4.1.4) in the regular supermarket chains. That indicates that there is actually a big demand for sustainable products and extending the offer could, therefore, bear a competitive advantage.

Simply extending the portfolio is not enough. Sustainable food manufacturers also have to rethink their communication and marketing & sales strategy. It is not possible to give specific recommendations as even within the food sector, different products may have very different purchase drivers, target customers, etc. In any case, manufacturers should consider the fact that often the consumer behavior is habitual, respectively the purchase decisions are made before the actual shopping trip (chapter 4.1.5). Referring to customer journey literature, the pre-purchasing phase (Lemon & Verhoef, 2016) gains importance and should be especially considered. It is essential to understand and address the factors influencing this decision, in order to get people to change their habits or even form them in favor of the respective company. Exploring the factors that influence the formation or breaking is a major challenge for these companies or could even be a direction for further research.

6.1.2. Public Policy

Apart from profit-driven firms, the findings are also valuable for public policy. Based on the findings of this thesis, necessary steps could be initiated in order to foster sustainable food consumption and make an effort in order to close the ABG.

As described in chapter 5.2.1, participants showed a low willingness to pay the price premium for sustainable products despite the fact that they could actually afford to pay it. Therefore, a sensible action for public policy would be the implementation of a GHG based tax that would raise the prices of unsustainable products. Rising the prices of unsustainable products would lift the barrier of sustainable products being too expensive (relatively) and would foster sustainable food consumption. If other taxes were lowered in return, the overall tax load, respectively the tax revenue for the German state could be kept constant.

Another ecologically worthwhile step, would be the development and implementation of a transparent, universally valid, and for all food products mandatory sustainability label. That way, people would not need to have extensive knowledge about sustainability. If it was a simple and concise concept like e.g. a traffic light system, people could additionally minimize the time necessary to evaluate the sustainability of products to simple checking the traffic light label with one look.

Moreover, sustainability and especially sustainable consumption should be part of the curriculum of any school type. It would guarantee that all people in Germany have a basic understanding of sustainability and would be able to evaluate different product alternatives regarding their sustainability.

6.1.3. Consumers

This thesis also bears implications for private consumers. They should actively reflect on and challenge their grocery shopping behavior regarding a possible ABG. Also, they should rethink and possibly adjust their habits as well as actively trying to take more time for grocery shopping and make conscious decisions (when grocery shopping). In this way, the gap could be closed or at least reduced, even if in special cases they give in to their desire and give themselves a treat.

If someone lacks the knowledge regarding sustainable food consumption, there are consumer protection and information apps like "Codecheck", which can help to evaluate the sustainability of products in an easy, fast and comprehensive way.

6.2. Theoretical Contribution

This thesis contributes to the consumer research on the ABG, as it is one of few studies focusing on sustainable food consumption in Germany and provides several explanation approaches for the gap. The work was built on previous research on the ABG, which was either generic (e.g. Prothero et al. (2011)) or had a focus on other industries (e.g. Schäufele and Hamm (2018)) and/or other countries (e.g. de Barcellos et al. (2011)). It is remarkable that most studies chose a rather quantitative approach to explore the ABG (e.g. Farjam et al. (2019), Götze and Naderer (2019), or Schäufele and Hamm (2018)). This thesis provided rich and new, qualitative insights as only a few studies had a comparable research object and setup (Aschemann-Witzel & Niebuhr Aagaard, 2014; Wiederhold & Martinez, 2018). Some of the findings of this thesis are conform to these studies (as described in chapter 2.2.3) but also new explanations were uncovered. Most important is the emergence of habit as main driver of consumer behavior regarding food consumption.

Wiederhold and Martinez (2018) also mentioned habit but described it mostly as the choice of the store consumers shop in, while the findings on habit in this thesis refer to the

product choices and. Also, the explanation approach that people try to minimize the time spent for grocery shopping is new.

Additionally, this thesis adds to behavioral research by applying the TPB and successfully extending it by one determinant, the self-efficacy. The extension worked out well, by clearly distinguishing between internal and external obstacles or reasons that keep people from putting a respective behavior into practice. Especially when it comes to removing these obstacles, this distinction is reasonable. While the food industry and public policy possibly have an influence on the external factors, it is more the call for the consumers themselves if it comes to the internal obstacles. With this extension the TPB was improved as one major criticism of the theory could be refuted.

At the same time, directions for further research can be derived from this thesis. One possibility could be quantifying and thereby verifying the qualitative findings of this study. A conceivable approach could be an online survey quantifying the attitude, subjective norm, PBC, self-efficacy, and the behavior itself and testing for possible interrelations.

Another sensible research direction could be to examine the habit formation in more detail in order to elaborate influencing factors, how people can change their habits themselves or how third parties, like public policy or food manufacturers or supermarket chains, can influence peoples' habits.

Last, it is important to mention that a generally valid and holistic measure of sustainability would highly benefit the whole research stream on sustainability and public policy by making sustainability tangible and comparable.

References

Ajzen, I. (1985). From intentions to actions: A theory of planned behavior [Book Section]. In *Action control: From cognition to behavior* (p. 11-39). Berlin, Heidelberg: Springer-Verlag.

Ajzen, I. (1988). *Attitudes, personality, and behavior* (US Edition ed.) [Book]. Stony Stratford, Milton Keynes, UK: Open University Educational Enterprise Limited.

Ajzen, I. (1991). The theory of planned behavior [Journal Article]. *Organizational Behavior and Human Decision Processes*, *50*(2), 179-211.

Ajzen, I. (2002). Perceived behavioral control, self-efficacy, locus of control, and the theory of planned behavior 1 [Journal Article]. *Journal of Applied Social Psychology*, *32*(4), 665-683.

Ajzen, I. (2005). *Attitudes, personality and behaviour* [Book]. McGraw-Hill Education (UK).

Ajzen, I., & Madden, T. J. (1986). Prediction of goal-directed behavior: Attitudes, intentions, and perceived behavioral control [Journal Article]. *Journal of Experimental Social Psychology*, *22*(5), 453-474.

Amatulli, C., De Angelis, M., Peluso, A. M., Soscia, I., & Guido, G. (2019). The effect of negative message framing on green consumption: An investigation of the role of shame [Journal Article]. *Journal of Business Ethics*, *157*(4), 1111-1132.

Antonetti, P., & Maklan, S. (2015). How categorisation shapes the attitude–behaviour gap in responsible consumption [Journal Article]. *International Journal of Market Research*, *57*(1), 51-72.

Armitage, C. J., & Conner, M. (2001). Efficacy of the theory of planned behaviour: A meta-analytic review [Journal Article]. *British Journal of Social Psychology*, *40*(4), 471-499.

Aschemann-Witzel, J., & Niebuhr Aagaard, E. M. (2014). Elaborating on the attitude–behaviour gap regarding organic products: young danish consumers and in-store food choice [Journal Article]. *International Journal of Consumer Studies*, *38*(5), 550-558.

Astrøm, A. N., & Rise, J. (2001). Young adults' intention to eat healthy food: Extending the theory of planned behaviour [Journal Article]. *Psychology and Health*, *16*(2), 223-237.

Atapattu, S. (2019). From "our common future" to sustainable development goals: Evolution of sustainable development under international law [Journal Article]. *Wisconsin International Law Journal*, *36*(2), 215-246.

Atteslander, P. (2013). *Verzerrungen im interview: zu einer fehlertheorie der befragung* (Vol. 32) [Book]. Springer-Verlag.

Babutsidze, Z., & Chai, A. (2018). Look at me saving the planet! the imitation of visible green behavior and its impact on the climate value-action gap [Journal Article]. *Ecological Economics*, *146*, 290-303.

Baker, S. (1996). Sustainable development and consumption: The ambiguities [Journal Article]. *The Oslo Ministerial Roundtable Conference on Sustainable Production and Consumption, Oslo, 6–10 February 1995*.

Bandura, A. (1986). Social foundations of thought and action [Journal Article]. *Englewood Cliffs, NJ*, *1986*.

Bandura, A., & Walters, R. H. (1977). *Social learning theory* (Vol. 1) [Book]. Prentice-hall Englewood Cliffs, NJ.

Bedard, S. A. N., & Tolmie, C. R. (2018). Millennials' green consumption behaviour: Exploring the role of social media [Journal Article]. *Corporate Social Responsibility and Environmental Management*, *25*(6), 1388-1396.

Bellmann, L., & Koch, T. (2019). *Ökologische Nachhaltigkeit in deutschen Unternehmen: Empirische Ergebnisse auf Basis des IAB-Betriebspanels 2018* [Webpage]. Retrieved from http://doku.iab.de/forschungsbericht/2019/fb0819.pdf

Bernyte, S. (2018). Sustainability marketing communications based on consumer values and principles [Journal Article]. *Regional Formation and Development Studies*, *26*(3), 26-35.

Biology Dictionary. (n.d.). *Biology Dictionary - Consumer* [Webpage]. Retrieved from https://biologydictionary.net/consumer/

Black, J., Hashimzade, N., & Myles, G. (2012). *A dictionary of economics* [Book]. OUP Oxford.

BMU. (2016). *Konsum und Ernährung* [Webpage]. Retrieved from https://www.bmu.de/themen/wirtschaft-produkte-ressourcen-tourismus/produkte-und-konsum/produktbereiche/konsum-und-ernaehrung/

Bohnsack, R., Meuser, M., & Geimer, A. (2018). *Hauptbegriffe qualitativer sozialforschung* [Book]. Opladen, Germany Toronto, Canada: UTB GmbH.

Boyon, N. (2020). *Climate Change and Consumer Behavior Global changes in consumer behavior in response to climate change* [Webpage]. Retrieved from https://www.ipsos.com/sites/default/files/ct/news/documents/2020-01/global-advisor-climate-change-consumer-behavior.pdf

Brühl, J. (n.d.). *Generationenübersicht* [Webpage]. Retrieved from https://den-wandel-gestalten.de/generationenubersicht/

Brough, A. R., Wilkie, J. E., Ma, J., Isaac, M. S., & Gal, D. (2016). Is eco-friendly unmanly? the green-feminine stereotype and its effect on sustainable consumption [Journal Article]. *Journal of Consumer Research*, *43*(4), 567-582.

Brundtland, G. H., Khalid, M., Agnelli, S., Al-Athel, S., & Chidzero, B. (1987). Our common future [Journal Article]. *New York*.

Bucic, T., Harris, J., & Arli, D. (2012). Ethical consumers among the millennials: A cross-national study [Journal Article]. *Journal of Business Ethics*, *110*(1), 113-131.

Bui, M., Droms, C. M., & Craciun, G. (2014). The impact of attitudinal ambivalence on weight loss decisions: Consequences and mitigating factors [Journal Article]. *Journal of Consumer Behaviour*, *13*(4), 303-315.

Bulut, Z. A., Kökalan Çımrin, F., & Doğan, O. (2017). Gender, generation and sustainable consumption: Exploring the behaviour of consumers from izmir, turkey [Journal Article]. *International Journal of Consumer Studies*, *41*(6), 597-604.

Bundesanstalt für Landwirtschaft und Ernährung. (2015). *Jeder Deutsche isst im Jahr run 60 kg Fleisch* [Webpage]. Retrieved from https://www.ble.de/SharedDocs/Pressemitteilungen/DE/2015/150928-Fleisch.html

Bundesministerium für Wirtschaft und Energie. (2019). *BMWi - Lebensmittelindustrie* [Webpage]. Retrieved from https://www.bmwi.de/Redaktion/DE/Artikel/Branchenfokus/Industrie/branchenfokus-lebensmittelindustrie.html

Bundesregierung. (2019). *Die Deutsche Nachhaltigkeitsstrategie* [Webpage]. Retrieved from https://www.bundesregierung.de/breg-de/themen/nachhaltigkeitspolitik/eine-strategie-begleitet-uns/die-deutsche-nachhaltigkeitsstrategie

Bundesverband der deutschen Ernährungsindustrie. (2019). *BVE - Statistikbroschüre* [Webpage]. Retrieved from https://www.bve-online.de/presse/infothek/publikationen-jahresbericht/bve-statistikbroschuere2019

Cambridge Dictionary. (2014). *Cambridge Dictionary: "Consumption"* [Webpage]. Retrieved from https://dictionary.cambridge.org/de/worterbuch/englisch/consumption

Carfora, V., Cavallo, C., Caso, D., Del Giudice, T., De Devitiis, B., Viscecchia, R., ... Cicia, G. (2019). Explaining consumer purchase behavior for organic milk: Including trust and green self-identity within the theory of planned behavior [Journal Article]. *Food Quality and Preference*, *76*, 1-9.

Carrigan, M., Szmigin, I., & Wright, J. (2004). Shopping for a better world? an interpretive study of the potential for ethical consumption within the older market [Journal Article]. *Journal of Consumer Marketing*, *21*(6), 401-417.

Carrington, M. J., Zwick, D., & Neville, B. (2016). The ideology of the ethical consumption gap [Journal Article]. *Marketing Theory*, *16*(1), 21-38.

Center for Generational Kinetics. (2016). *An intro to generations* [Webpage]. Retrieved from https://genhq.com/faq-info-about-generations/

Chan, D. K., & Fishbein, M. (1993). Determinants of college women's intentions to tell their partners to use condoms 1 [Journal Article]. *Journal of Applied Social Psychology*, *23*(18), 1455-1470.

Chen, Y., Ghosh, M., Liu, Y., & Zhao, L. (2019). Media coverage of climate change and sustainable product consumption: Evidence from the hybrid vehicle market [Journal Article]. *Journal of Marketing Research*, *56*(6), 995-1011.

Chung, S.-S., & Leung, M. M.-Y. (2007). The value-action gap in waste recycling: the case of undergraduates in hong kong [Journal Article]. *Environmental Management*, *40*(4), 603-612.

Collins Cobuild. (n.d.). *Collins Dictionary: " Consumption"* [Web-

page]. Retrieved from https://www.collinsdictionary.com/de/worterbuch/englisch/consumption

Connolly, J., & Shaw, D. (2006). Identifying fair trade in consumption choice [Journal Article]. *Journal of Strategic Marketing*, *14*(4), 353-368.

Dean, M., Raats, M. M., & Shepherd, R. (2012). The role of self-identity, past behavior, and their interaction in predicting intention to purchase fresh and processed organic food [Journal Article]. *Journal of Applied Social Psychology*, *42*(3), 669-688.

de Barcellos, M. D., Krystallis, A., de Melo Saab, M. S., Kügler, J. O., & Grunert, K. G. (2011). Investigating the gap between citizens' sustainability attitudes and food purchasing behaviour: empirical evidence from brazilian pork consumers [Journal Article]. *International Journal of Consumer Studies*, *35*(4), 391-402.

Dickinson, J. L. (2009). The people paradox: Self-esteem striving, immortality ideologies, and human response to climate change [Journal Article]. *Ecology and Society*, *14*(1).

D'Orlando, F., & Sanfilippo, E. (2010). Behavioral foundations for the keynesian consumption function [Journal Article]. *Journal of Economic Psychology*, *31*(6), 1035-1046.

Dunning, D. (2007). Self-image motives and consumer behavior: How sacrosanct self-beliefs sway preferences in the marketplace [Journal Article]. *Journal of Consumer Psychology*, *17*(4), 237–249.

Dutta, S. (2014). Components of ecological behaviour and consumer attitude towards sustainable green environment: An empirical analysis on kolkata citizens [Journal Article]. *Globsyn Management Journal*, *8*(1/2), 53-63.

eKitchen. (2017). *Nachhaltig einkaufen: Was stammt aus meiner Region?* [Webpage]. Retrieved from https://www.ekitchen.de/kuechengeraete/kuehlschrank/ratgeber/regionale-produkte-181113.html

Ellen, P. S., Wiener, J. L., & Cobb-Walgren, C. (1991). The role of perceived consumer effectiveness in motivating environmentally conscious behaviors [Journal Article]. *Journal of Public Policy & Marketing*, *10*(2), 102-117.

Emblem, A. (2012). *Packaging technology: Fundamentals, materials and processes* [book]. Elsevier.

Ericsson, K. A. (2006). Protocol analysis and expert thought: Concurrent verbalizations of thinking during experts' performance on representative tasks [Journal Article]. *The Cambridge Handbook of Expertise and Expert Performance*, 223-241.

Ericsson, K. A., & Simon, H. A. (1985). *Protocol analysis: Verbal reports as data* [Book]. Cambridge, USA: The MIT Press.

Ernährung, B. f. L. u. (2019). *Fleisch und Geflügel* [Webpage]. Retrieved from https://www.ble.de/SharedDocs/Downloads/DE/BZL/Daten-Berichte/Fleisch/Fleisch_2018.xlsx?__blob=publicationFile&v=2

Farjam, M., Nikolaychuk, O., & Bravo, G. (2019). Experimental evidence of an environmental attitude-behavior gap in high-cost situations [Journal Article]. *Ecological Economics*, *166*, 106434.

Ferguson, K. E., Hair Jr, J. F., Silva, R. V., Oliveira-Brochado, A., Mollah, M. M. H., & Musharuf, M. (2017). Consumer perceptions of sustainability: an exploratory study [Journal Article]. *International Journal of Business, Marketing, & Decision Science*, *10*(1), 1-21.

Fishbein, M. (1967). A consideration of beliefs and their role in attitude measurement [Journal Article]. *Readings in Attitude Theory and Measurement*, 257-266.

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, and behavior: An introduction to theory and research* [Book]. Reading, Massachusetts: Addison-Wesley Publishing Company, Inc.

Fishbein, M., & Ajzen, I. (1980). *Understanding attitudes and predicting social behavior* [Book]. Englewood Cliffs, N.J.: Prentice-Hall, Inc.

Flatley, A. (2019). *Palmöl-Report: Nestlé bezieht Palmöl aus Kinderarbeit* [Webpage]. Retrieved from https://utopia.de/ratgeber/solidar-suisse-nestle-bezieht-palmoel-aus-kinderarbeit/

Flick, U. (1991). *Handbuch qualitative sozialforschung* [Book]. München: Psychologie Verlags Union.

Flick, U. (2011). Triangulation [Book Section]. In (3rd ed.). Wiesbaden: Springer Fachmedien Wiesbaden GmbH.

Flick, U. (2018). *An introduction to qualitative research* (6th ed.) [Book]. London, UK: Sage Publications Ltd.

Flynn, R., Bellaby, P., & Ricci, M. (2009). The 'value-action gap'in public attitudes towards sustainable energy: the case of hydrogen energy [Journal Article]. *The Sociological Review*, *57*(2), 159-180.

Frank, P., & Brock, C. (2018). Bridging the intention–behavior gap among organic grocery customers: The crucial role of point-of-sale information [Journal Article]. *Psychology & Marketing*, *35*(8), 586-602.

Gabler, C. B., Butler, T. D., & Adams, F. G. (2013). The environmental belief-behaviour gap: Exploring barriers to green consumerism [Journal Article]. *Journal of Customer Behaviour*, *12*(2-3), 159-176.

Gaes, G. G., Kalle, R. J., & Tedeschi, J. T. (1978). Impression management in the forced compliance situation: Two studies using the bogus pipeline [Journal Article]. *Journal of Experimental Social Psychology*, *14*(5), 493–510.

Geddes, B. (1990). How the cases you choose affect the answers you get: Selection bias in comparative politics [Journal Article]. *Political Analysis*, *2*, 131–150.

Gentina, E., Butori, R., Rose, G. M., & Bakir, A. (2014). How national culture impacts teenage shopping behavior: Comparing french and american consumers [Journal Article]. *Journal of Business Research*, *67*(4), 464-470.

Giri, A., & Biswas, D. (2019). *Research methodology for social sciences* [Book]. New Delhi, India: SAGE Publications India Pvt Ltd.

Global Carbon Project. (2019). *Entwicklung des weltweiten C02-Ausstoßes in den Jahren 1995 - 2018* [Webpage]. Retrieved from http://www.globalcarbonatlas.org/en/CO2-emissions

Gómez-Luciano, C. A., de Aguiar, L. K., Vriesekoop, F., & Urbano, B. (2019). Consumers' willingness to purchase three alternatives to meat proteins in the united kingdom, spain, brazil and the dominican republic [Journal Article]. *Food Quality and Preference*, *78*(31), 103732.

Gollwitzer, P. M. (1999). Implementation intentions: strong effects of simple plans. [Journal Article]. *American Psychologist*, *54*(7), 493.

Govind, R., Singh, J. J., Garg, N., & D'Silva, S. (2019). Not walking the walk: How dual attitudes influence behavioral outcomes in ethical consumption [Journal Article]. *Journal of Business Ethics*, *155*(4), 1195-1214.

Grieco, C. (2018). What do social entrepreneurs need to walk their talk? understanding the attitude–behavior gap in social impact assessment practice [Journal Article]. *Nonprofit Management and Leadership*, *29*(1), 105-122.

Grünberg, J., Nieberg, H., & Schmidt, T. (2010). Treibhausgasbilanzierung von lebensmitteln (carbon footprints): überblick und kritische reflektion [Journal Article]. *Landbauforschung–vTI Agriculture and Forestry Research*, *60*(2), 53-72.

Götze, S., & Naderer, G. (2019). From the attitude-behaviour-gap to a producer-people-gap? [Journal Article]. *International Journal of Sustainable Development & World Ecology*, *26*(5), 377-388.

Gummerus, J., Liljander, V., & Sihlman, R. (2017). Do ethical social media communities pay off? an exploratory study of the ability of facebook ethical communities to strengthen consumers' ethical consumption behavior [Journal Article]. *Journal of Business Ethics*, *144*(3), 449-465.

Gupta, S., & Agrawal, R. (2018). Environmentally responsible consumption: Construct definition, scale development, and validation [Journal Article]. *Corporate Social Responsibility and Environmental Management*, *25*(4), 523-536.

Halkier, B. (2001). Routinisation or reflexivity? consumers and normative claims for environmental consideration [Book Section]. In A. W. Jukka Gronow (Ed.), *Ordinary consumption* (p. 25 - 44). London: Routledge.

Hameed, I., Waris, I., & ul Haq, M. A. (2019). Predicting eco-conscious consumer behavior using theory of planned behavior in pakistan [Journal Article]. *Environmental Science and Pollution Research*, *26*(15), 15535-15547.

Han, T.-I., & Stoel, L. (2017). Explaining socially responsible consumer behavior: A meta-analytic review of theory of planned behavior [Journal Article]. *Journal of International Consumer Marketing*, *29*(2), 91-103.

Handelsblatt. (2013). *Wem Nestlé das Wasser abgräbt* [Webpage]. Retrieved from https://www.handelsblatt.com/unternehmen/handel-konsumgueter/lebensmittelkonzern-schwere-anschuldigungen-gegen-nestle/7782074-2.html?ticket=

ST-27883780-sWMnQ9hD4AKJ3W2rdH5V-ap4

Hanss, D., Böhm, G., Doran, R., & Homburg, A. (2016). Sustainable consumption of groceries: The importance of believing that one can contribute to sustainable development [Journal Article]. *Sustainable Development*, *24*(6), 357-370.

Hiller, A., & Woodall, T. (2019). Everything flows: A pragmatist perspective of trade-offs and value in ethical consumption [Journal Article]. *Journal of Business Ethics*, *157*(4), 893-912.

Hák, T., Janoušková, S., & Moldan, B. (2016). Sustainable development goals: A need for relevant indicators [Journal Article]. *Ecological Indicators*, *60*, 565-573.

Holdershaw, J., Gendall, P., & Wright, M. (2011). Predicting blood donation behaviour: further application of the theory of planned behaviour [Journal Article]. *Journal of Social Marketing*.

Johnstone, M.-L., & Hooper, S. (2016). Social influence and green consumption behaviour: A need for greater government involvement [Journal Article]. *Journal of Marketing Management*, *32*(9-10), 827-855.

Johnstone, M.-L., & Tan, L. P. (2015). Exploring the gap between consumers' green rhetoric and purchasing behaviour [Journal Article]. *Journal of Business Ethics*, *132*(2), 311-328.

Jonas, K., & Doll, J. (1996). Eine kritische Bewertung der Theorie überlegten Handelns und der Theorie geplanten Verhaltens [Journal Article]. *Zeitschrift für sozialpsychologie*, *27*(1), 18–31.

Jones, K. (1996). Trust as an affective attitude [Journal Article]. *Ethics*, *107*(1), 4-25.

Jung, H. J., Kim, H., & Oh, K. W. (2016). Green leather for ethical consumers in china and korea: Facilitating ethical consumption with value–belief–attitude logic [Journal Article]. *Journal of Business Ethics*, *135*(3), 483-502.

Juvan, E., & Dolnicar, S. (2014). The attitude–behaviour gap in sustainable tourism [Journal Article]. *Annals of Tourism Research*, *48*, 76-95.

Kacen, J. J., & Lee, J. A. (2002). The influence of culture on consumer impulsive buying behavior [Journal Article]. *Journal of Consumer Psychology*, *12*(2), 163-176.

Kaiser, F. G., Wölfing, S., & Fuhrer, U. (1999). Environmental attitude and ecological behaviour [Journal Article]. *Journal of Environmental Psychology*, *19*(1), 1-19.

Kersten, P., McCambridge, A., M. Kayes, N., Theadom, A., & McPherson, K. M. (2015). Bridging the gap between goal intentions and actions: a systematic review in patient populations [Journal Article]. *Disability and Rehabilitation*, *37*(7), 563-570.

Kim, E., Ham, S., Yang, I. S., & Choi, J. G. (2013). The roles of attitude, subjective norm, and perceived behavioral control in the formation of consumers' behavioral intentions to read menu labels in the restaurant industry [Journal Article]. *International Journal of Hospitality Management*, *35*, 203-213.

Kim, S.-Y., Yeo, J., Sohn, S. H., Rha, J.-Y., Choi, S., Choi, A.-y., & Shin, S. (2012). Toward a composite measure of green consumption: an exploratory study using a korean sample [Journal Article]. *Journal of Family and Economic Issues*, *33*(2), 199-214.

Klöckner, C. A., & Blöbaum, A. (2010). A comprehensive action determination model: Toward a broader understanding of ecological behaviour using the example of travel mode choice [Journal Article]. *Journal of Environmental Psychology*, *30*(4), 574–586.

Kohlbacher, F. (2006). The use of qualitative content analysis in case study research [Conference Proceedings]. In *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research* (Vol. 7, pp. 1–30).

Kumar, J. V., & Chamola, P. (2019). Exploring antecedents of responsible consumption using structural equation modeling [Journal Article]. *Serbian Journal of Management*, *14*(1), 77-95.

Kuo, F., & Young, M. (2008). A study of the intention–action gap in knowledge sharing practices [Journal Article]. *Journal of the American Society for Information Science and Technology*, *59*(8), 1224-1237.

Lapierre, M. A., & Rozendaal, E. (2018). Cross-national differences in consumer socialization, development, and behavior: a comparison of children in the united states and the netherlands [Journal Article]. *Journal of Children and Media*, *12*(4), 388-405.

Lee, H., & Cheon, H. (2018). Exploring korean consumers' attitudes toward ethical consumption behavior in the light of affect and cognition [Journal Article]. *Journal of International Consumer Marketing*, *30*(2), 98-114.

Lee, K. (2008). Opportunities for green marketing: young consumers [Journal Article]. *Marketing Intelligence & Planning*, *26*(6), 573-586.

Lemon, K. N., & Verhoef, P. C. (2016). Understanding customer experience throughout the customer journey [Journal Article]. *Journal of Marketing*, *80*(6), 69–96.

Liu, H., Wang, L., & Koehler, M. J. (2019). Exploring the intention-behavior gap in the technology acceptance model: A mixed-methods study in the context of foreign-language teaching in china [Journal Article]. *British Journal of Educational Technology*, *50*(5), 2536-2556.

Luchs, M. G., Phipps, M., & Hill, T. (2015). Exploring consumer responsibility for sustainable consumption [Journal Article]. *Journal of Marketing Management*, *31*(13-14), 1449-1471.

Mack, B., Tampe-Mai, K., Kouros, J., Roth, F., Taube, O., & Diesch, E. (2019). Bridging the electricity saving intention-behavior gap: A german field experiment with a smart meter website [Journal Article]. *Energy Research & Social Science*, *53*, 34-46.

Madden, T. J., Ellen, P. S., & Ajzen, I. (1992). A comparison of the theory of planned behavior and the theory of reasoned action [Journal Article]. *Personality and Social Psychology Bulletin*, *18*(1), 3-9.

Maksan, M. T., Kovačić, D., & Cerjak, M. (2019). The influence of consumer ethnocentrism on purchase of domestic wine: Application of the extended theory of planned behaviour [Journal Article]. *Appetite*, *142*, 104393.

Manrai, A. K. (2018). New research on cross-cultural and cross-national comparisons in advertising and consumer behavior [Journal Article]. , *31*(1), 1-3.

Mayring, P. (2010). Qualitative inhaltsanalyse. handbuch qualitative forschung in der psychologie [Incollection]. In *VS Verlag fur Sozialwissenschaften* (pp. 601–613).

Mayring, P. (2016). *Einführung in die qualitative sozialforschung* [Book]. Beltz.

McCrindle, M., & Wolfinger, E. (2014). *The abc of xyz: Understanding the global generations* (3rd ed.) [Book]. Bella Vista, Australia: McCrindle Research Pty Ltd.

Meredith, G., & Schewe, C. (1994). The power of cohorts [Journal Article]. *American Demographics*, *16*(2), 22-27, 30-31.

Miller, R. K., & Washington, K. (2019). *Consumer behavior 2017-2018* (book No. 1). Richard K Miller & Associates.

Minton, E. A., Jeffrey Xie, H., Gurel-Atay, E., & Kahle, L. R. (2018). Greening up because of god: The relations among religion, sustainable consumption and subjective well-being [Journal Article]. *International Journal of Consumer Studies*, *42*(6), 655-663.

Moderne-Landwirtschaft. (2016). *Regional und nachhaltig - was heißt das?* [Webpage]. Retrieved from https://www.moderne-landwirtschaft.de/regional-und-nachhaltig-was-heisst-das

Moore, R., & Burrus, J. (2019). Predicting stem major and career intentions with the theory of planned behavior [Journal Article]. *The Career Development Quarterly*, *67*(2), 139-155.

Moraes, C., Carrigan, M., & Szmigin, I. (2012). The coherence of inconsistencies: Attitude–behaviour gaps and new consumption communities [Journal Article]. *Journal of Marketing Management*, *28*(1-2), 103-128.

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review [Journal Article]. *European Journal of Social Psychology*, *15*(3), 263–280.

Nguyen, H. V., Nguyen, C. H., & Hoang, T. T. B. (2019). Green consumption: Closing the intention-behavior gap [Journal Article]. *Sustainable Development*, *27*(1), 118-129.

Osterburg, B., Nieberg, H., Röder, N., Isermeyer, F., Haenel, H.-D., Hahne, J., ... others (2009). *Erfassung, Bewertung und Minderung von Treibhausgasemissionen des deutschen Agrar-und Ernährungssektors: Studie im Auftrag des Bundesministeriums für Ernährung, Landwirtschaft und Verbraucherschutz* (Technical Report). Arbeitsberichte aus der vTI-AgrarÖkonomie. Retrieved from https://www.econstor.eu/handle/10419/39359

Ouellette, J. A., & Wood, W. (1998). Habit and intention in everyday life: The multiple processes by which past behavior predicts future behavior. [Journal Article]. *Psychological Bulletin*, *124*(1), 54.

Parkinson, J., David, P., & Rundle-Thiele, S. (2017). Self-efficacy or perceived behavioural control: Which influences consumers' physical

activity and healthful eating behaviour maintenance? [Journal Article]. *Journal of Consumer Behaviour*, *16*(5), 413-423.

Pelau, C. (2011). Analysis of consumer behavior for different product groups [Journal Article]. *Management & Marketing*, *6*, 101.

Penker, M., & Elmar, S. (2015). *Nachhaltigkeit durch Regionalität? Pro und Contra* [Webpage]. Retrieved from https://www.forum-wirtschaftsethik.de/nachhaltigkeit-durch-regionalitaet-pro-und-contra/

Perera, C., Auger, P., & Klein, J. (2018). Green consumption practices among young environmentalists: a practice theory perspective [Journal Article]. *Journal of Business Ethics*, *152*(3), 843-864.

PETA. (2019). *Die schlimmsten Lebensmittelskandale in Deutschland: Fipronil, BSE,...* [Webpage]. Retrieved from https://www.peta.de/skandalchronik

Pfeffer, I., Englert, C., & Mueller-Alcazar, A. (2019). Perceived stress and trait self-control interact with the intention–behavior gap in physical activity behavior [Journal Article]. *Sport, Exercise, and Performance Psychology*.

Priporas, C.-V., Stylos, N., & Fotiadis, A. K. (2017). Generation z consumers' expectations of interactions in smart retailing: A future agenda [Journal Article]. *Computers in Human Behavior*, *77*, 374-381.

Procter, L., Angus, D. J., Blaszczynski, A., & Gainsbury, S. M. (2019). Understanding use of consumer protection tools among internet gambling customers: Utility of the theory of planned behavior and theory of reasoned action [Journal Article]. *Addictive Behaviors*, *99*, 106050.

Prothero, A., Dobscha, S., Freund, J., Kilbourne, W. E., Luchs, M. G., Ozanne, L. K., & Thøgersen, J. (2011). Sustainable consumption: Opportunities for consumer research and public policy [Journal Article]. *Journal of Public Policy & Marketing*, *30*(1), 31-38.

Raithel, J. (2008). *Quantitative Forschung* [Book]. Springer.

Ramayah, T., Lee, J. W. C., & Lim, S. (2012). Sustaining the environment through recycling: An empirical study [Journal Article]. *Journal of Environmental Management*, *102*, 141-147.

Rana, N. P., Slade, E., Kitching, S., & Dwivedi, Y. K. (2019). The it way of loafing in class: Extending the theory of planned behavior (tpb) to understand students' cyberslacking intentions [Journal Article]. *Computers in Human Behavior*, *101*, 114-123.

Rasool, S., Shakur, A., ani Mohammad, M., Mughal, Y. H., & Awang, Z. (2019). Validating a measure for altruistic self towards the responsible plate food consumption: A mix method approach [Journal Article]. *International Journal of Business & Society*, *20*(1), 211-228.

Redondo, I., & Puelles, M. (2017). The connection between environmental attitude–behavior gap and other individual inconsistencies: a call for strengthening self-control [Journal Article]. *International Research in Geographical and Environmental Education*, *26*(2), 107-120.

Reicks, M., Smith, C., Henry, H., Reimer, K., Atwell, J., & Thomas, R. (2003). Use of the think aloud method to examine fruit and vegetable purchasing behaviors among low-income african american women [Journal Article]. *Journal of Nutrition Education and Behavior*, *35*(3), 154-160.

Rhodes, R. E., & Courneya, K. S. (2003). Investigating multiple components of attitude, subjective norm, and perceived control: An examination of the theory of planned behaviour in the exercise domain [Journal Article]. *British Journal of Social Psychology*, *42*(1), 129-146.

Robb, J., Haggar, J., Lamboll, R., & Castellanos, E. (2019). Exploring the value–action gap through shared values, capabilities and deforestation behaviours in guatemala [Journal Article]. *Environmental Conservation*, *46*(3), 226-233.

Robinot, l., Ertz, M., & Durif, F. (2017). Jingle bells or 'green'bells? the impact of socially responsible consumption principles upon consumer behaviour at christmas time [Journal Article]. *International Journal of Consumer Studies*, *41*(6), 605-617.

Rossmann, C. (2011). *Theory of reasoned action-theory of planned behavior* [Book]. Baden-Baden: Nomos Verlagsgesellschaft mbH & Co. KG.

Saarela, A., Kantanen, T. T., Lapveteläinen, A. T., Mykkänen, H. M., Karppinen, H. A., & Rissanen, R. L. (2013). Combining verbal analysis protocol and wireless audiovisual observation to examine consumers' supermarket shopping behaviour [Journal Article]. *International Journal of Consumer Studies*, *37*(5), 577-584.

Sachs, J. D. (2012). From millennium development goals to sustainable development goals [Journal Article]. *The Lancet*, *379*(9832), 2206-2211.

Saddawi-Konefka, D., Schumacher, D. J., Baker, K. H., Charnin, J. E., & Gollwitzer, P. M. (2016). Changing physician behavior with implementation intentions: closing the gap between intentions and actions [Journal Article]. *Academic Medicine*, *91*(9), 1211-1216.

Salciuviene, L., Buenaventura, V. E. C., & Lee, K. (2019). Employee proactiveness to engage in sustainable consumption leading to societal benefits [Journal Article]. *Engineering Economics*, *30*(1), 112-120.

Schifter, D. E., & Ajzen, I. (1985). Intention, perceived control, and weight loss: an application of the theory of planned behavior [Journal Article]. *Journal of Personality and Social Psychology*, *49*(3), 843.

Schmidt, K. (2019). Predicting the consumption of expired food by an extended theory of planned behavior [Journal Article]. *Food Quality and Preference*, *78*, 103746.

Schäufele, I., & Hamm, U. (2018). Organic wine purchase behaviour in germany: Exploring the attitude-behaviour-gap with data from a household panel [Journal Article]. *Food Quality and Preference*, *63*, 1-11.

Scott, K. A., & Weaver, S. T. (2018). The intersection of sustainable consumption and anticonsumption: Repurposing to extend product life spans [Journal Article]. *Journal of Public Policy & Marketing*, *37*(2), 291-305.

Shaw, D., McMaster, R., & Newholm, T. (2016). Care and commitment in ethical consumption: An exploration of the 'attitude–behaviour gap' [Journal Article]. *Journal of Business Ethics*, *136*(2), 251-265.

Shepherd, R., & Raats, M. M. (1996). Attitudes and Beliefs in Food Habits [Book Section]. In *Food choice, acceptance and consumption* (p. 346-364). Springer.

Shin, Y. H., Im, J., Jung, S. E., & Severt, K. (2018). The theory of planned behavior and the norm activation model approach to consumer behavior regarding organic menus [Journal Article]. *International Journal of Hospitality Management*, *69*, 21-29.

Shobeiri, S., Rajaobelina, L., Durif, F., & Boivin, C. (2016). Experiential motivations of socially responsible consumption [Journal Article]. *International Journal of Market Research*, *58*(1), 119-139.

Song, S. Y., & Kim, Y.-K. (2018). Theory of virtue ethics: do consumers' good traits predict their socially responsible consumption? [Journal Article]. *Journal of Business Ethics*, *152*(4), 1159-1175.

Sonneveld, K., James, K., Fitzpatrick, L., & Lewis, H. (2005). Sustainable packaging: how do we define and measure it [Conference Proceedings]. In *22nd IAPRI Symposium* (pp. 1–9).

Southgate, D. (2017). The emergence of generation z and its impact in advertising: Long-term implications for media planning and creative development [Journal Article]. *Journal of Advertising Research*.

Sparks, P., Guthrie, C. A., & Shepherd, R. (1997). The dimensional structure of the perceived behavioral control construct 1 [Journal Article]. *Journal of Applied Social Psychology*, *27*(5), 418-438.

Sparks, P., Hedderley, D., & Shepherd, R. (1992). An investigation into the relationship between perceived control, attitude variability and the consumption of two common foods [Journal Article]. *European Journal of Social Psychology*, *22*(1), 55-71.

Sparks, P., & Shepherd, R. (1992). Self-identity and the theory of planned behavior: Assesing the role of identification with" green consumerism" [Journal Article]. *Social Psychology Quarterly*, 388-399.

Statistisches Bundesamt. (2019). *Fortschreibung des Bevölkerungsstandes Deutschland* [Webpage]. Retrieved from https://www-genesis.destatis.de/genesis/online/data;sid=DA14875F972C0D8F4DC4B8DA4F232BC9.GO_1_2?operation=abruftabelleBearbeiten&levelindex=2&levelid=1571135640079&auswahloperation=abruftabelleAuspraegungAuswaehlen&auswahlverzeichnis=ordnungsstruktur&auswahlziel=werteabruf&selectionname=12411-0005&auswahltext=&werteabruf=Werteabruf

Stern, P. C. (1997). Toward a working definition of consumption for environmental research and policy [Journal Article]. *Environmentally Significant Consumption: Research Directions*, 12-35.

Stern, P. C. (2000). New environmental theories: toward a coherent theory of environmentally significant behavior [Journal Article]. *Journal of Social Issues*, *56*(3), 407-424.

Strauss, A. L. (1998). *Grundlagen qualitativer Sozialforschung: Datenanalyse und Theoriebildung in der empirischen soziologischen Forschung*

(Vol. 2) [Book]. München: Wilhelm Fink Verlag.

Sustainable Development Knowledge Platform. (n.d.). *Sustainable consumption and production* [Webpage]. Retrieved from https://sustainabledevelopment.un.org/topics/sustainableconsumptionandproduction#

Tam, K.-P., & Chan, H.-W. (2018). Generalized trust narrows the gap between environmental concern and pro-environmental behavior: Multilevel evidence [Journal Article]. *Global Environmental Change*, *48*, 182-194.

Tegut. (2015). *Was ist Nachhaltigkeit?* [Webpage]. Retrieved from https://www.tegut.com/aktuell/artikel/was-ist-nachhaltigkeit.html

Testa, F., Sarti, S., & Frey, M. (2019). Are green consumers really green? exploring the factors behind the actual consumption of organic food products [Journal Article]. *Business Strategy and the Environment*, *28*(2), 327-338.

The Nielson Company. (2015). *The Sustainability Imperative* [Webpage]. Retrieved from https://www.nielsen.com/us/en/insights/report/2015/the-sustainability-imperative-2/

Trudel, R. (2019). Sustainable consumer behavior [Journal Article]. *Consumer Psychology Review*, *2*(1), 85-96.

Tukker, A., Huppes, G., Guinée, J., Heijungs, R., de Koning, A., van Oers, L., ... Jansen, B. (2006). Environmental impact of products (eipro) [Journal Article]. *Analysis of the Life Cycle Environmental Impacts Related to the Final Consumption of the EU-25. Ispra*.

Łuczka, W., & Smoluk-Sikorska, J. (2017). Sustainable consumption - between theory and practice [Conference Proceedings]. In *International Scientific Conference Rural Development 2017* (p. 1161-1166).

Umweltbundesamt. (2013). *Nahrungsmittelindustrie* [Webpage]. Retrieved from https://www.umweltbundesamt.de/themen/wirtschaft-konsum/industriebranchen/nahrungs-futtermittelindustrie-tierhaltungsanlagen/nahrungsmittelindustrie#textpart-1

Umweltbundesamt. (2017). *Gelebte Nachaltigkeit: von der Nische in den Mainstream* [Webpage]. Retrieved from https://www.umweltbundesamt.de/gelebte-nachaltigkeit-von-der-nische-in-den#textpart-1

Umweltbundesamt. (2019a). *Umweltbewusstsein in Deutschland* [Webpage]. Retrieved from https://www.umweltbundesamt.de/themen/nachhaltigkeit-strategien-internationales/gesellschaft-erfolgreich-veraendern/umweltbewusstsein-in-deutschland

Umweltbundesamt. (2019b). *Umweltbewusstsein und Umweltverhalten* [Webpage]. Retrieved from https://www.umweltbundesamt.de/daten/private-haushalte-konsum/umweltbewusstsein-umweltverhalten#textpart-2

Umweltbundesamt. (2020). "*Grüne*" *Produkte: Marktzahlen* [Webpage]. Retrieved from https://www.umweltbundesamt.de/daten/private-haushalte-konsum/konsum-produkte/gruene-produkte-marktzahlen#umsatz-mit-grunen-produkten

United Nations. (n.d.-a). *Goal 12: Ensure sustainable consumption and production patterns* [Webpage]. Retrieved from https://sustainabledevelopment.un.org/post2015/transformingourworld

United Nations. (n.d.-b). *Sustainable Development Communication Materials* [Webpage]. Retrieved from https://www.un.org/sustainabledevelopment/news/communications-material/

United Nations. (2015). *Transforming our world: the 2030 Agenda for Sustainable Development* [Webpage]. Retrieved from https://sustainabledevelopment.un.org/post2015/transformingourworld

United Nations Environment Programme. (2015). *Sustainable consumption and production policies* [Webpage]. Retrieved from https://www.unenvironment.org/explore-topics/resource-efficiency/what-we-do/sustainable-consumption-and-production-policies

Van der Merwe, K., & Maree, T. (2016). The behavioural intentions of specialty coffee consumers in south africa [Journal Article]. *International Journal of Consumer Studies*, *40*(4), 501-508.

Verma, S., Petersen, A. C., & Lansford, J. E. (2019). *Working Toward Sustainable Development for All* [Journal Article]. Hogrefe Publishing.

Wei, C., Chiang, C., Kou, T., & Lee, B. C. (2017). Toward sustainable livelihoods: Investigating the drivers of purchase behavior for green products [Journal Article]. *Business Strategy and the Environment*, *26*(5), 626-639.

Wey Smola, K., & Sutton, C. D. (2002). Generational differences: Revisiting generational work values for the new millennium [Journal Article]. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, *23*(4), 363-382.

White, K., Habib, R., & Dahl, D. W. (2020). A review and framework for thinking about the drivers of prosocial consumer behavior [Journal Article]. *Journal of the Association for Consumer Research*, *5*(1), 2-18.

White, K., Habib, R., & Hardisty, D. J. (2019). How to shift consumer behaviors to be more sustainable: A literature review and guiding framework [Journal Article]. *Journal of Marketing*, *83*(3), 22-49.

Wiederhold, M., & Martinez, L. F. (2018). Ethical consumer behaviour in germany: The attitude-behaviour gap in the green apparel industry [Journal Article]. *International Journal of Consumer Studies*, *42*(4), 419-429.

Williams, L. T., Germov, J., Fuller, S., & Freij, M. (2015). A taste of ethical consumption at a slow food festival [Journal Article]. *Appetite*, *91*, 321-328.

Wirth, W., Von Pape, T., & Karnowski, V. (2008). An integrative model of mobile phone appropriation [Journal Article]. *Journal of Computer-Mediated Communication*, *13*(3), 593–617.

World Population Review. (2019). *GDP Ranked by Country 2020* [Webpage]. Retrieved from http://worldpopulationreview.com/countries/countries-by-gdp/

World-Resource-Institute. (2016). *CAIT - Country Greenhouse Gas Emissions Data* [Webpage]. Retrieved from http://datasets.wri.org/dataset/cait-country

Xu, H., You, X., & Liu, Y. (2019). Tourists' socially responsible consumption: Concept and scale development [Journal Article]. *Social Behavior and Personality: An International Journal*, *47*(11), 1-15.

Yap, S., & Lee, C. K. C. (2013). Does personality matter in exercise participation? [Journal Article]. *Journal of Consumer Behaviour*, *12*(5), 401-411.

Young, W., Hwang, K., McDonald, S., & Oates, C. J. (2010). Sustainable consumption: green consumer behaviour when purchasing products [Journal Article]. *Sustainable Development*, *18*(1), 20-31.

Junior Management Science 6(3) (2021) 468-488

Junior Management Science

journal homepage: www.jums.academy

The Role of the European Central Bank in a Sustainable Financial System

Nele Braun

University of Applied Sciences Darmstadt

Abstract

The Paris Agreement acknowledged climate change as an urgent threat to the planet and human society. To fulfil the aim of limiting global warming, public and private investments and especially long-term investments are supposed to shift towards sustainable practices. Given the high investments required to pursue a sustainable financial system, it will be essential to involve the financial sector, as well as its participants and authorities. This thesis discussed the role the European Central Bank (ECB) could play in a transition towards a sustainable financial system. First, the framework conditions and the need for a sustainable financial system were explained, in particular the recent developments like for instance the introduction of the EU Taxonomy regulation, a classification system for sustainable activities which aims to provide clarity and limit the risk of green washing. After that, it was outlined how climate-related risks can spread to the financial system and why central banks are concerned of them. Three links for the relation between climate change and the financial system were identified – physical risks, transition risks and liability risks. In particular, the impact of climate change on price stability, financial stability and the portfolio management of central banks were examined. The objectives and the strategy of the ECB were described, to establish a base for the subsequent analysis of their instruments. Furthermore, the European Green Deal, an answer of the European Union to the challenges caused by climate change was presented.

Keywords: Sustainable finance; green finance; central banking; sustainability.

1. Introduction

In January 2020, the World Meteorological Organization (WMO) acknowledged the year 2019 as the second warmest year ever recorded. The WMO Statement on the State of Global Climate 2019 documented the physical symptoms of climate change as well as the impact on socio-economic development. In the context of this statement, UN chief António Guterres emphasized that the world is far away from reaching the 2°C target as determined in the Paris Agreement (see United Nations, 2020, p. 1).

In 2015, the European Union and 196 states adopted the Paris Agreement to establish the target of limiting global warming below 2°C above pre-industrial levels. In 2016, the countries have set themselves the goal of limiting the global temperature increase to 1.5°C (see IPCC, 2018, p. 8-10). The Paris Agreement acknowledged climate change as an urgent threat to the planet and human society. It aims "(...) to increase the ability of countries to deal with the impacts of climate change, and at making finance flows consistent with a low GHG emissions and climate-resilient pathway." (United Nations Framework Convention on Climate Change , 2020, p. 1). With this objective, the Paris Agreement sends a clear signal to the financial markets that a profound change in global investment flows is required. To fulfil this aim, public and private investments and especially long-term investments are supposed to shift towards sustainable practices (see Thwaites, Whitley, Wright, & Ott, 2018, p. 1).

Thence climate change poses a challenge to financial institutions. In January 2019, the European Parliament appealed to the European Central Bank to consider environmental, social and governance (ESG) factors in its policies and to revise its program for the purchase of corporate sector securities "(...) to better support environmentally sustainable initiatives" (European Parliament, 2020, p. 1). The primary mandate of the European Central Bank is to preserve price stability and an inflation target of below, but close to, 2%. The Treaty on the Functioning of the European Union emphasizes the primary objective on price stability, but also states "(...) Without prejudice to the objective of price stability, the ESCB shall support the general economic policies

DOI: https://doi.org/10.5282/jums/v6i3pp468-488

in the Union with a view to contributing to the achievement of the objectives of the Union(...)" (Estella, 2018, p. 78).

Consequently, a debate has arisen as to whether the European Central Bank should take the aspect of sustainability into account in its policies. This thesis will therefore approach the research question: "What role can the ECB fulfil in a sustainable financial system?".

First, the framework conditions and the need for a sustainable financial system will be explained. Afterwards, the effects of climate change on the financial system and particularly on central banks are evaluated. After a short introduction of the European Green Deal, the objectives and strategy of the ECB will be described to establish a base for the subsequent analysis. Different instruments in the areas of the ECB will be analysed with regards of the objective to support a transition towards a sustainable financial system. Subsequently the restrictions of the ECB to contribute be discussed. Eventually a qualitative approach to value the introduced instruments will be drawn and applied to two instruments.

2. A sustainable financial system

The term "sustainable" is defined as the quality of causing little or no damage to the environment and therefore able to continue for a long time (see Cambridge Dictionary, 2020, p. 1).

A "financial system" is described as a complex interaction of markets and institutional units, with the purpose of mobilizing funds for investment and providing facilities for the financing of commercial activities (see International Monetary Fund, 2004, p. 12).

At this point in time, there is no single definition of a "sustainable financial system". According to the European Commission, the term refers to the consideration of environmental and social considerations in investment decisions. The inclusion of this criteria leads to more investments in long-term and sustainable investments (see European Commission, 2018b, p. 2). The environmental considerations are specifically about adapting to climate change and mitigating its consequences. They also refer to general environmental aspects, like carbon dioxide pollution. Social considerations can relate to issues like inequality and employment and investment in people and communities (see European Commission, 2018b, p. 2). In 2018, the European Commission published an action plan for a more sustainable financial system, addressed to the major EU institutions, including the European Central Bank. The action plan named "Financing sustainable growth" emphasizes that, to meet the targets of the Paris Agreement, major public and private investments are required to transform the economy of the EU (see European Commission, 2019a, p. 1). The European Commission Vice-President for the Euro and Social Dialogue, Valdis Dombrovskis, argues that "(...) Europe needs between € 175 to € 290 billion in additional yearly investment in the next decades. We want a quarter of the EU budget to contribute to climate action as of 2021. Yet, public money will not be enough (...) the EU has proposed hard law to incentivise private capital to flow to green projects." (European Commission, 2019a, p. 1). In line with the action plan, the European Commission established a Technical Expert Group (TEG) on sustainable finance (see European Commission, 2018a, p. 1). The TEG mainly consists of members from the academic; public; finance and civil society sector. Their mandate is to assist the European Commission, inter alia, in the development of an EU classification system (see European Commission, 2018a, p. 1). The first section of the plan aims for an EU classification system for sustainable activities. A uniform classification system within the EU should provide clarity, which economic activities can be considered as sustainable. The European Commission followed this approach with a proposal for a taxonomy regulation. A taxonomy is defined as a uniform procedure, with which objects are classified according to certain criteria into categories or classes (see Cambridge Dictionary, 2020, p. 1). In December 2019, the European Parliament and the Council arranged a political agreement on the taxonomy (see Council of the European Union, 2019, p. 2). The TEG on sustainable finance developed recommendations for relevant criteria of the Taxonomy Regulation, culminating in a final report, published in March 2020. The aim of the EU Taxonomy is to serve as a tool to help financial market participants to identify environmentally friendly investments (see EU Technical Expert Group on Sustainable Finance, 2020, p. 2-7). The EU Taxonomy specifies performance thresholds for economic activities which "(...) make a substantive contribution to one of six environmental objectives (...) do no significant harm (DNSH) to the other five, where relevant; meet minimum safeguards (e.g., OECD Guidelines on Multinational Enterprises (...)" (EU Technical Expert Group on Sustainable Finance, 2020, p. 2). The six environmental objectives consist of climate change mitigation; climate change adaptation; sustainable and protection of water and marine resources; transition to a circular economy; pollution prevention and control as well as the protection and restoration of biodiversity and ecosystems (see EU Technical Expert Group on Sustainable Finance, 2020, p. 19-25).

Furthermore, the Taxonomy Regulation introduces detailed new disclosure requirements for two different groups. All companies belonging under the scope of the Non-Financial Reporting Directive (NFRD) are already required to publish regular reports on the environmental and social impact of their activities. Additionally, they will have to reveal the extent to which they are aligned with the Taxonomy Regulation (see EU Technical Expert Group on Sustainable Finance, 2020, p. 27). Companies under the scope of the NFRD are approximately 6000 large public-interest companies each with more than 500 employees, including listed companies, banks and insurance companies (see European Commission, 2017, p. 1). All financial market participants offering financial products in the EU are obligated to make Taxonomy disclosures, with detailed information about the sustainability of underlying investments and their proportion of the whole investment, depending on the specific financial product (see EU Technical Expert Group on Sustainable Finance,

2020, p. 37).

In conclusion, there may is no single or legal definition for a sustainable finance system yet, but the European Commission considers a more sustainable financial system a key figure in the transition process to a low-emission, more resource-efficient circular economy and seeks to reform the financial system accordingly (see European Commission, 2018b, p. 1-4). The first step was taken with the development of the Taxonomy Regulation, which will ensure a uniform definition of sustainability and classification of sustainability in the financial system. The taxonomy for climate change mitigation and climate change adaptation is scheduled for 2020, its full application by the end of 2021. For the other objectives, the taxonomy should be set up by the end of 2021, the application is scheduled for the end of 2022 (see European Council, 2020: 1).

2.1. The link between climate change and the financial system

In the following section, the connection of climate change and the financial system will be examined in more detail. It will outline how the consequences of climate change can spread to the financial system and its participants like commercial banks; households; companies and insurers.

In 2015, climate change-related risks on the financial systems came to the fore after the Governor of the Bank of England, Mark Carney, delivered a speech named "The tragedy of the horizons". In his speech, Mark Carney defined three channels through which climate change can affect the financial system, particularly financial stability: physical risks, liability risks and transition risks (see Carney, 2015, p. 5-7).

Physical risks impact the society directly and have the potential to affect the economy: because of climate change, the world experiences extreme weather events like flooding, droughts and storms (see Bank of England, 2015, p. 4). Physical risks can spread to the financial system through the disruption of global supply chains, as well as through the damage of property of companies. An example therefore are the Thailand floods in 2011, adding up to an economic damage of US$45 billion and US$12 billion in insurance claims as well as the closing of over 10.000 manufacturing companies due to property damage (see Bank of England, 2015, p. 29). If the possible impacts from physical risks are insured, this has a direct effect on the risk management behaviour of insurance companies. If the demand for weather-related insurance rises, insurance companies need to increase fees to have a sufficient pay-out buffer (see Bank of England, 2015, p. 28). Andrew Howard, Head of Sustainable Research at Schroders Investment Management, identified oil & gas, utilities and basic resources as the sectors most exposed to the physical impact of climate change (see Howard & Hassler, 2018, p. 1). According to their study, the cost of insuring physical assets corresponds to more than 3% of their market values (see Howard & Hassler, 2018: 1). Apart from financial insurance companies, physical risks also have an impact on commercial banks. Another physical risk example would be the increase of credit losses due to an unpredictable, extreme weather event. Climate change can lead to increased credit risks for banks, therefore experts argue that banks should integrate climate-related risks into their risk management framework (see Diakatos & Bisisidis, 2020, p. 1). Institutional investors with large portfolios usually manage their risks through a balanced portfolio construction, including diversification, liquidity and asset selection (see Gründl, Gal, & Dong, 2016, p. 8). Physical risks pose a challenge on some of these investors since their well-established portfolio risk management strategies do not take into account the possible physical risks from climate change (see Benedetti et al. 2019: 21). All in all, for the financial system physical risks can be felt especially through consequences along the value chain for business customers, counterparties and equity holdings.

Liability risks are the risks that could emerge every day by parties who suffered consequences from climate change and consequently pursue compensation for their losses (see Carney, 2015, p. 6). The possible addressees can be environmentally harmful companies, but also insurance companies. A recent example is the Rhode Island's climate liability suit. The state of Rhode Island accuses 21 companies, who knowingly contributed to climate change and failed to reasonably inform Rhode Island citizens about the risks posed by their products (see Savage, 2019, p. 1). This example illustrates, that firms are also exposed to legal risks through climate change. Measuring liability risks is considered a major challenge because of the uncertainty and the variations in the different legal frameworks of different countries (see Bolton, Despres, Pereira da Silva, Samama, & Svartzman, 2020, p. 41). The Prudential Regulation Authority (PRA), a financial services regulatory body from the Bank of England, considers liability risk most relevant for general insures, because of the possibility of an increase of third-party liability claims (see Bank of England, 2015, p. 7). According to the European Central Bank, insurance companies are particularly important for a stable financial system, because they are large investors in financial markets and are often closely linked to banks, thus are likely to increase systemic risk (see European Central Bank, 2009, p. 160).

Transition risks refer to the financial risks that can occur when moving to a less polluting, low carbon economy (see Carney, 2015, p. 6). These include risks from regulatory changes and technological changes, who are implemented to pursue the target of decarbonisation (see Ferrer, 2019, p. 35). Examples for regulatory changes are strict emission limits as proposed in the Paris Agreement. Risks from technological changes refer to the speed of new technologies and the disruption of whole sectors that come with them, that may result in higher costs (see Carney, 2015, p. 11). Moreover, there is a risk of changes in demand and services in all sectors resulting from emerging consumer preferences, as the society becomes more sensitive to environmental and sustainability issues (see Ferrer, 2019, p. 35). For instance, if the policies of the government adopt the regulations of the Paris Agreement, a large proportion of fossil fuel reserves could not be used anymore. This would ultimately lead to changes in the

value of investments owned by banks and insurance companies in sectors like oil and gas (see Bank of England, 2017, p. 1). The Figure 1 below provides a summarizing overview of the mentioned risks and their potential impacts on the financial system.

After explaining the link between climate change and the financial system, the next section discusses the impact of climate change on central banks.

2.2. Climate change and central banks

Central banks implement the monetary policy and control the money supply of their nation, usually commissioned with the maintenance of a low inflation target. The current operations of central banks are often subject to frequent criticism, however their role within the economy is universally accepted (see Goodhart, 1995, p. 60). After the financial crisis in 2008, central banks worldwide injected more than USD12 trillion of extra cash into the financial system to propel economic recovery (see Goodman & McCormick, 2019, p. 1). Given the before mentioned immense investments required to swift towards a sustainable financial system, the public discussion arose whether central banks should occupy a central role in this challenge. This abstract briefly reviews recent discussions about the roles of central banks and their responsibilities.

2.2.1. The role of central banks

According to Mishkin (2005, p. 411) the six main goals of central banks and their monetary policy are high employment; economic growth; price stability; interest-rate stability; stability of financial markets, and stability in foreign exchange markets.

They can be divided into three main objectives: the maintenance of price stability, the maintenance of financial stability and the support of wider economic policy objectives (including economic growth and high employment). The financial crisis in 2008 sparked off a discussion about central banks and their inflation targeting strategy, contributing to the goal of price stability. Most economists believed that the best contribution central banks could make, was to focus on low and stable inflation rates (see Volz, 2017, p. 6). This assumption was underlined by the inflation targeting framework. The framework is characterized by the announcement of official target ranges for the inflation rate, fostering the acceptance of low inflation as the main goal of monetary policy (see Bernanke et al. 1997: 97). Following the financial crisis, financial stability shifted into focus. Thus, the inflation targeting framework has been criticized for "(...) failing to address concerns other than price stability and most importantly for its disregard of financial stability." (Volz, 2017, p. 6). Even when environmental objectives are not included by name in a central banks mandate, the inclusion of sustainability aspects may be considered to carry out the main goals, price stability and financial stability. In the following abstract, these arguments will be outlined.

2.2.2. Climate change and price stability

Climate change can affect price stability (see Batten, Sowerbutts, & Tanaka, 2016, p. 23). Physical and transition risks can influence the economy and the inflation performance. Climate-related shocks, like droughts and floods, can have an impact on monetary policy through supply-side and demand-side shocks (see Bolton et al., 2020, p. 16). Demand shocks are unpredictable events that increase or decrease the demands for goods and services, supply shocks increase or decrease the supply or cost of goods and services (see McKibbin, Morris, Panton, & Wilcoxen, 2017, p. 1). Climate change related extreme weather events can affect agriculture production and consequently food prices (see Volz, 2017, p. 9). A study from the European Central Bank states that natural disasters, particularly storms, can trigger a prompt increase in food price inflation for the first six months (see Parker, 2016, p. 25). Climate change policies like a carbon tax can affect energy production and energy prices (see McKibbin et al., 2017, p. 1-4). Both prices are components of inflation, and there is evidence that both prices are highly volatile (see Parker, 2017, p. 28).

These price changes can have indirect effects on core inflation (see Volz, 2017, p. 9). Core inflation is defined as the change in prices of goods and services without those from food and energy sectors (see Amadeo, 2019, p. 1).

Therefore, Volz (2017, p. 9) argues that "Factors driving food and energy prices thus need to be included in central banks' long-term inflation outlook analysis".

2.2.3. Climate change and financial stability

There are two main, before mentioned transmission channels through which climate change can affect financial stability.

First, financial stability can be impaired through the effects of extreme weather events and natural disasters, referred to as physical risks as explained in abstract 2.1. Secondly, the financial system can be impacted by the "(...) the uncertainties related to the timing and speed of the process of adjustment towards a low-carbon economy, including the impact of the related policy action and potentially disruptive technological progress on the asset prices of carbon-intensive sectors", referred to as transition risks (Giuzio et al., 2019, p. 1). Physical risks refer to the economic costs and financial losses caused by extreme weather events and the effects of long-term changes in the environment (see Bolton et al., 2020, p. 17). Through damages to assets, capital depreciation may take place, resulting in abrupt repricing of whole asset classes. As outlined in abstract 2.1, weather-related losses can affect the solvency of firms, insurance companies and households, and consequently the whole financial system.

Transition risks are the risks that are associated with mitigation challenges. They include policy and legislation changes, reputational impacts, technological limitations as well as shifts in market preferences (see Bolton et al., 2020, p. 18). To reach the climate targets of the Paris Agreement, a major part of fossil fuel reserves is required to stay in the

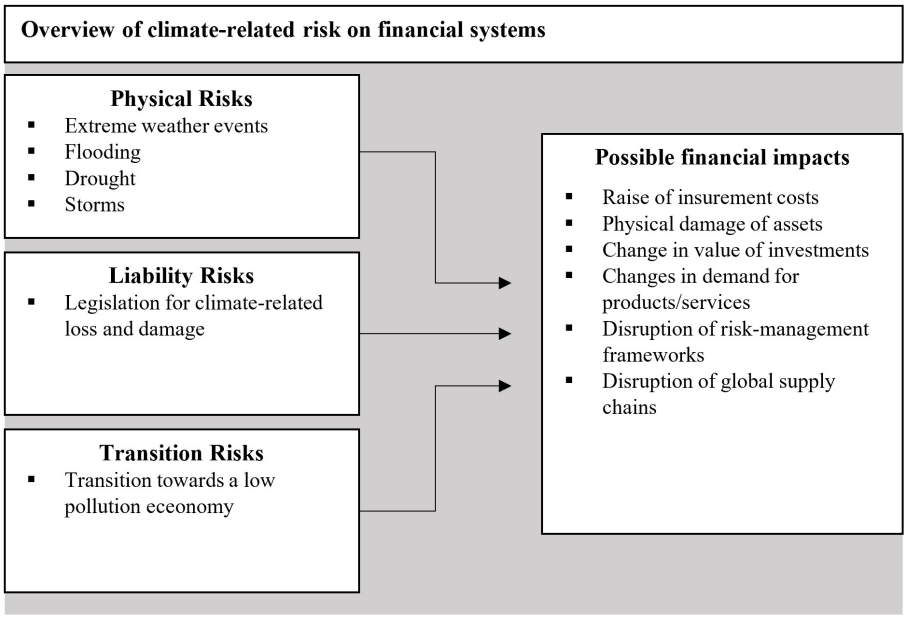


Figure 1: Overview of climate-related risk on the financial system (Own illustration)

ground (see Verkuijl, Piggot, Lazarus, van Asselt, & Erickson, 2018, p. 1-3). If these fossil reserves are not extracted, so-called stranded assets can arise (see Ivlevia, 2017: 1). Stranded assets refer to assets which "(...) suffer from unanticipated or premature write-offs, downward revaluations or are converted to liabilities (...)" (Ansar, Caldecott, & Tilbury, 2013, p. 9). These stranded assets pose a risk on companies that rely on fossil fuels, since these assets contain the risk of experiencing a devaluation or depreciation, that is not reflected in the value of the company (see Bolton et al., 2020, p. 19). Consequently, possible new regulations and legislation with the aim of decarbonizing will significantly impact carbon businesses.

A report by the Carbon Disclosure Project (CDP) states that out of 859 analysed companies, 75% consider transition risks as a major factor influencing their business in the future (see CDP, 2019, p. 1).

Carney (2016, p. 7) argues that an abrupt movement towards a low-carbon economy could significantly influence financial stability by destabilizing markets, as costs and challenges that come with regulations require a reassessment of asset values. In conclusion, the main risk for financial stability from climate change is a widespread fall in financial asset prices that could impact the financial system (see Fisher & Alexander, 2019, p. 16). Therefore, Volz (2017, p. 20) argues that these risks need to be addressed in the financial stability and macro prudential policy frameworks of central banks, at least to the extent that central banks have the task of maintaining financial stability.

2.2.4. Climate change and portfolio management

As outlined before, the effects of physical and transition risks may require reassessments of financial assets. Thus, climate change will also affect the investment portfolios under central banks' management. Investment portfolios under the central banks' vary depending on their mandates, but they may include policy portfolios; own portfolios; pension portfolios and third-party portfolios (see NGFS, 2019a, p. 7). Therefore Cecot (2019, p. 2) argues that central banks need to consider environmental, social and governance (ESG) risks in their portfolio management. Central banks are committed to consider the demands of third-parties for whom they may manage assets (see Cecot, 2019, p. 2). As public institutions, they "(...) would face reputational risk if they fail to address stakeholders' concerns related to climate change and if they don't lead by example." (Cecot, 2019, p. 2). To address this concern, the Network for Greening the Financial System (NGFS) developed a sustainable investment guide for central banks' portfolio management (see NGFS, 2019a, p. 1-21). The Network for Greening the Financial System is a global network of central banks and supervisory authorities. It was established in 2017 with the aim to contribute to the transition towards a sustainable, low-carbon economy and to analyse consequences for the financial system related to climate change. In April 2019, the NGFS released "A call for action", acknowledging climate change as a financial risk and providing recommendations for central banks, supervisors, policy makers and financial institutions to manage climate-related risks (see NGFS, 2019a, p. 1-4).

Apart from the financial system, calls for further climate action resulted in the European Green Deal, a concept from the European Commission, introduced in December 2019 (see European Commission, 2019b, p. 2). The main targets and areas of the European Green Deal will be outlined in the next abstracts.

2.3. The European Green Deal

The European Green Deal was presented by the European Commission by Ursula von der Leyen in December 2019. Its main aim is to reduce the net emissions of greenhouse gases in the European Union (EU) to zero by 2050, making it the first continent to become climate neutral (see European Commission, 2019b, p. 2). The European Green Deal belongs to the six priorities of the European Commission from 2019-2024. In the following abstract, the areas, measures and targets of the concept will be outlined.

2.3.1. Main aspects of the European Green Deal

"We are determined to tackle climate change and turn it into an opportunity for the European Union.", concludes Ursula von der Leyen, president of the European Commission, after introducing the European Green Deal, also referred to as the "roadmap" towards green transition, to the European Council in December 2019 (see European Commission, 2019b, p. 1).

The European Green Deal is a response to the climate challenges posed on the EU. It aims to transform the EU's economy into a resource-efficient with zero net emissions of greenhouse gases by 2050. To reach this target, in March 2020 the European Commission proposed the European Climate Law to the European Parliament, the Council, the Economic and Social Committee and the Committee of the Regions. The European Climate Law contains a legally binding target of zero greenhouse gases by 2050 and measures to control the progress towards it (see European Commission, 2020a, p. 1-4). The Green Deal aims to divorce the growth of the economy from the use of resources (see European Commission, 2020a, p. 2-4). The natural capital of the EU should be safeguarded, as well as its citizens, from climate change related risks (see European Commission, 2019b, p. 22). The term "natural capital" refers to natural resources like air; water; soil; minerals; forests and plants (see Cantrell, 2020, p. 1). The Green Deal states that the transition that comes with it, needs to put people first, just like everyone who faces major challenges coming with it. Moreover, a close cooperation of citizens and EU authorities and institutions is required to enable a fundamental transition (see European Commission, 2019b, p. 20-22). A huge amount of public investment will be required to carry out the transition. The creation of a sustainable, financial system which facilitates sustainable investments is considered as essential. Further, the European Green Deal states that the EU can use its competences to encourage neighbouring countries to pursue sustainability goals and, where appropriate, form joint alliances (see European Commission, 2019b, p. 20-22). The main purpose here is to establish global standards for sustainability and to integrate climate issues into international relations (see European Commission, 2019b, p. 2; 9).

The Green Deal contains various policy areas for actions. The area of sustainable food systems pursues a sustainable agriculture in the EU by modernising agriculture with innovative technologies and following the Common Agriculture Policy (CAP), which supports farmers (see Stolze, Sanders, Kasperczyk, Madsen, & Meredith, 2016, p. 9). The policy area of clean energy focuses on the decarbonisation of the EU's energy system inter alia by integrating renewable energy sources (see European Commission, 2019b, p. 6). The area "Sustainable Industry" includes a new circular action plan that aims to modernise the economy of the EU, especially tackling resource intense sectors like steel and cement (see European Commission, 2019b, p. 7). Further areas include the protection of biodiversity; the elimination of pollution as well as the promotion of sustainable mobility (see European Commission, 2019b, p. 4-10). In May 2020, two core strategies of the Green Deal were introduced. The "Farm to Fork Strategy" is designed to make the food system of the EU more sustainable, inter alia by reducing the use of chemical pesticides and by a revision of the date marking in order to minimize food waste (see European Commission, 2020b, p. 1). The "EU Biodiversity Strategy for 2030" is intended to protect the biodiversity of the European Union, planned measures include the establishing of more protected areas of land and sea, the planting of 3 billion trees until 2030 and a zero pollution action plan for air, soil and water (see European Commission, 2020b, p. 1).

The figure below provides an overview of the main aspects and measures in the different policy areas. It shows, that two of the key points of the Green Deal are the financing of the whole transition as well as the mantra "Leave no one behind", which aims to foster the message of an inclusive Europe. In addition to that, the figure shows that the framework of the Green Deal is based on the enhancement of innovation and research.

In January 2020, the European Commission proposed a € 100 billion "Just Transition Mechanism", which includes three pillars – the "Just Transition Fund" (JTF), a just transition scheme as well as a public sector loan facility (see European Commission, 2020c, p. 1).

The Just Transition Mechanism is an element of the € 1 trillion investment plan for the Green Deal. In May 2020, they published details regarding the public sector loan facility in cooperation with the European Investment Bank (EIB). The facility will include € 1.5 billion from the EU budget and around € 10 billion in loans from the EIB.

For the JTF, the European Commission proposed € 10 billion from the EU budget and € 30 billion from the "Next Generation EU" fund (see European Commission, 2020e, p. 1). The emergency instrument "Next Generation EU" is a temporary measure to support the reconstruction of the EU, it increased the EU budget by additional € 750 billion (see European Commission, 2020d, p. 1).

After the presentation of the European Green Deal by Ursula von der Leyen, the president of the European Central Bank, Christine Lagarde, announced that she wants the European Central Bank to contribute more actively towards the challenges of climate change. She stated that the role of sustainability will be subject during the strategy review of the European Central Bank in 2020 (see European Central Bank, 2020b, p. 11). As a result, the public discussion emerged, whether the European Central Bank should intervene in the

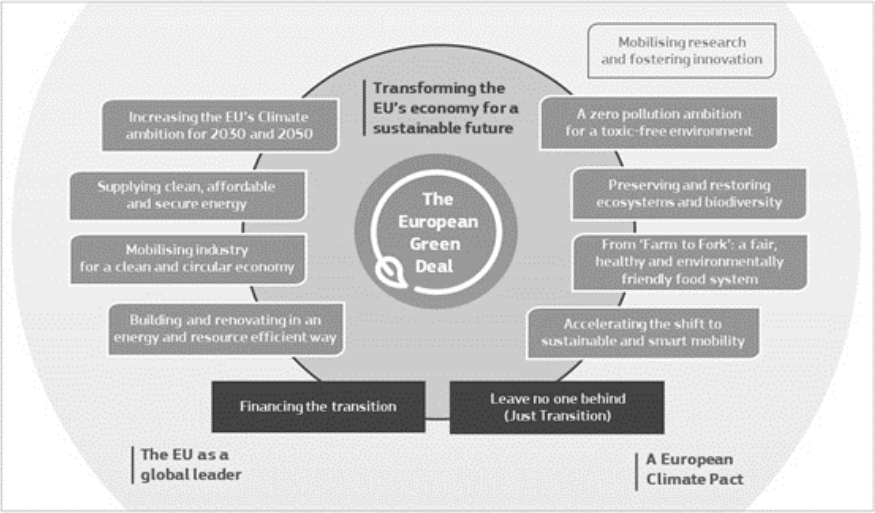


Figure 2: Overview of the European Green Deal (European Commission, 2019b)

fight against climate change or even include the topic of climate change in their strategy. To further investigate this topic, in the next abstract the objectives and the mandate of the European Central Bank will be outlined.

3. The possible roles of the ECB in a sustainable financial system

In the following, the possible roles of the ECB in a sustainable financial system will be discussed. Since there is widespread disagreement whether climate-related activities should be in scope of the ECB, the objectives, mandate and strategy of the ECB will be outlined first. Central banks can employ various policy instruments to pursue sustainability targets (see Volz, 2017, p. 20). The following section distinguishes four different policy areas, including central bank soft power, macro prudential regulation, micro prudential supervision and monetary policy. For each of the four areas, different policy instruments and tools, which could be introduced to support the transition towards a more sustainable financial system, will be discussed.

3.1. Objectives and mandate of the ECB

The Treaty on the Functioning of the European Union assigned objectives and a normative mandate to the ECB (see Thiele, 2013, p. 18). The Article 127 TFEU states that the primary objective of the European Central Bank is to maintain price stability (see European Union, 2012, p. 56). With the maintenance of price stability, the ECB can make the best contribution within their monetary policy to the growth of the economy and the creation of jobs (see European Central Bank, 2011, p. 7).

Price stability is not only the primary objective of the ECB's monetary policy, but also an objective of the whole European Union (see European Union, 2012, p. 303). Moreover, the ECB "(...) shall support the general economic policies in the Union (...) and act in accordance with the principle of an open market economy with free competition, favouring an efficient allocation of resources (...)" (European Union, 2012, p. 56). The Article 127 para. 2 TFEU defines four basic competences: definition and conduct of monetary policy of the Union, the operation of foreign exchange ports, the holding and management of the Member States' official foreign reserves and the promotion of the smooth operation of payment systems (see European Union, 2012, p. 56).

The Article 128 of the TFEU authorises the ECB to issue banknotes in the community. Banknotes issued by the ECB and national central banks are the sole banknotes having the exclusive status of legal tender (see European Union, 2012, p. 57). The TFEU also entrusts the ECB the responsibility for specific tasks concerning the prudential supervision of credit institutions. The ECB performs these tasks within a single supervisory mechanism (SSM) consisting of the ECB and the national competent authorities (see European Union, 2012, p. 56-58). To achieve its primary objective, the ECB pursues a monetary policy strategy, that will be outlined in the next abstract.

3.2. The strategy of the ECB

The monetary policy strategy is the general approach of the ECB to achieve its primary objective – the maintenance of price stability. The strategy establishes a comprehensive framework within which decisions are taken on the appropriate level of short-term interest rates (see European Central Bank, 2011, p. 55-63). The main aim of the strategy is to secure a consistent and systematic approach to monetary policy decisions. Since monetary policy should be "(...) forward-looking and preemptive, focusing on the medium term, firmly anchoring inflation expectations and be broadly based." (European Central Bank, 2011, p. 63), the ECB's strategy contains two core elements.

The first element is the quantitative definition of price stability. To enumerate the objective of price stability, the Governing Council of the ECB provided a quantitative definition in 1998: "Price stability shall be defined as a year-on-year increase in the Harmonised Index of Consumer Prices (HICP) for the euro area of below 2%. Price stability is to be maintained over the medium term." (Gerdesmeier, 2009, p. 8).

The HICP measures consumer price inflation in the member states of the European Union, Norway, Iceland and Switzerland by measuring the change over time in the prices of consumer goods and services, used or paid for by euro area households (see Eurostat, 2019, p. 1-3). In 2003, the Governing Council clarified that it aims to maintain inflation rates "below, but close to 2% over the medium term." (European Central Bank, 2011, p. 64). The quantitative definition of price stability aims to make the primary objective of the ECB more transparent and comprehensible. It also provides a measure for the public to make the ECB liable to account in the event of deviations from the target. The definition also serves as a guide for the public to form expectations about future price developments (see European Central Bank, 2011, p. 64).

The second element is a two-pillar approach which serves as a tool to organise and evaluate relevant information for the risk-assessment of price stability (see Gerdesmeier, 2009, p. 63). This approach consists of two analytical perspectives: the economic analysis and the monetary analysis.

The economic analysis serves to assess the short and medium-term determinants of price developments. It concentrates on the real activity and financial conditions in the economy from the perspective of the interplay between supply and demand in goods and services (see Gerdesmeier, 2009, p. 63).

The monetary analysis provides a medium to long-term perspective on risks to price stability. It focuses on the long-run link between money and prices and is used as a cross-checking tool for the indications from the economic analysis (see European Central Bank, 2011, p. 69-79).

In conclusion, the two-pillar approach aims to ensure that "all relevant information is used in the assessment of the risks to price stability and that appropriate attention is paid to different perspectives and the cross-checking of information in order to come to an overall judgement of the risks to price stability" (Gerdesmeier, 2009, p. 63). The strategy of the ECB in general also provides a guidance for explaining monetary policy decisions to the public in a transparent way. The figure below shows an overview of the stability-oriented monetary policy strategy of the ECB. On the fundament of a sufficient set of information, the two analyses are made and cross-checked against each other. Based on the result of the two analyses and the overall risks to price stability, the Governing Council adopts monetary policy decisions, which should contribute to the primary objective.

In January 2020, the ECB started its strategy review process which was extended until mid-2021 due to the Covid-19 pandemic. The last review of the strategy was in 2003, and since the world experienced fundamental changes in the economy since then, the ECB decided to review its strategy (see Lagarde, 2020, p. 1). The review will cover the definition of price stability; the inflation rate; the monetary policy instruments; the analysis process of the economy as well as topics related to employment and communication practices. Furthermore, it will deal with the topic of climate change and the effects of it on monetary policy (see Schnabel, 2020, p. 1; Lagarde, 2020, p. 1). Closely linked to the communication practices of the ECB is their soft power, therefore the next chapter will discuss the possible role of soft power of the ECB in a sustainable financial system.

3.3. The role of soft power of the ECB

Soft power refers to the use of power by political institutions, actors and states over the society and other states. This power is not based on military resources (see v. Von Voss-Wittig, 2006, p. 1-2). Soft power dispenses with coercive measures. It is based on the persuasive power of the actors to convince them of their interests. Its resources include the function to lead by example, attractiveness and the communication of own norms and values (see v. Von Voss-Wittig, 2006, p. 1-2). Joseph S. Nye defined soft power as "the power of attraction" (see Nye, 2017, p. 2).

As Chandavarkar (1987, p. 4) states "More than statutes and formal powers it is the central bank's status, expertise, and influence, in relation to the government (...) and the financial community, which defines its promotional opportunity-set as a regulator, innovator, participant, guarantor, and catalyst of financial development". Consequently, the ECB can highlight the importance of climate change challenges by soft power. It already engaged in promoting the discussion on green finance by dedicating an own chapter on their website to climate change and by publishing various speeches and research on the topic. With the membership in the Network for Greening the Financial System (NGFS) as well as the inclusion of the topic in their strategy review, the ECB signals that it is aware of the importance of the whole challenge. The next abstract will discuss a specific possible measure of soft power: the introduction of voluntary green finance guidelines and frameworks.

3.3.1. Voluntary green finance guidelines and frameworks

The aim of voluntary green finance guidelines is the enhancement of sustainable finance. There are various possibilities of their exact design. In this case, green finance guidelines with the aim of the promotion of green lending will be discussed. So far there are no uniform standards of green loans. The Loan Market Association (2018, p. 2) defines green loans as "(...) any type of loan instrument made available exclusively to finance or re-finance, in whole or in part, new and/or existing eligible Green Projects." . The LMA published a list of indicative categories for Green Projects, including renewable energy; pollution prevention; clean transportation and wastewater management (see Loan Market Association, 2018, p. 4). To support banks with a guidance towards green lending, green credit guidelines are required

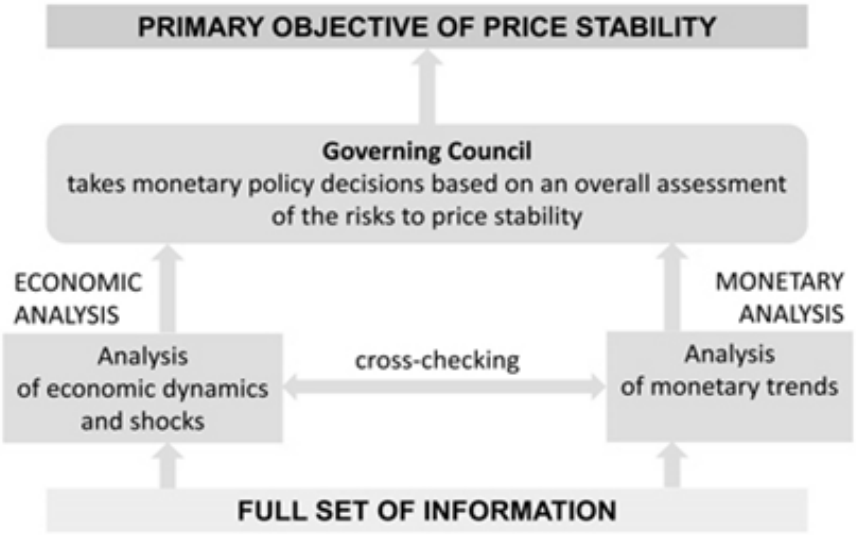


Figure 3: The stability oriented monetary policy strategy of the ECB (European Central Bank, 2020j, p. 1)

(see Volz, 2017, p. 16). One possibility for the ECB would be to develop and publish such voluntary guidelines to provide banks with guidance towards green lending.

A country that already introduced green credit guidelines is China. In 2012, the China Banking Regulatory Commission (CBRC) published the "Green Credit Guidelines" which aim to encourage Chinese banks to support low-carbon economies, manage environmental risks in the credit lending process and to foster sustainable performance in their operations (see Qingrong, Wang, Yiting, & Lina, 2013, p. 4-6).

The second chapter of the "Green Credit Guidelines" focuses on the organisation and management of banks and pleads that the board of directors should create and support green credit concepts (see China Banking Regulatory Commission, 2012, p. 1). It transfers the responsibility of defining exact roles within the green credit concepts to the senior management (see China Banking Regulatory Commission, 2012, p. 1). However, as of today the green credit policy of China did not fulfil its aims. Causes of the failure to comply with the guidelines are a lack of environmental information and unclear policy standards (see China Banking Regulatory Commission, 2012, p. 12). Furthermore, Volz (2015, p. 30) argues that clearly binding regulations which include compulsory components are required to cause a meaningful change in the lending behaviour of banks. However, voluntary principles offer the advantage of supporting the progress of green finance without the potential costs and delays which might occur with some binding regulations (see United Nations Environment Programme, 2017, p. 20).

A voluntary green guideline framework of the ECB could resemble like the following. The framework could include two stages with two different time horizons. The foundation of the first stage was already established by the publication of research and education from ECB, as well as the recognition of an absence of a standard definition of the terms "green" and "sustainable" (see Mersch, 2018, p. 1). In the first stage with a time horizon of 1 year, the ECB could deliver advice for banks and financial institutions in line with the Taxonomy regulation that will start applying in December 2021. The ECB could encourage banks to send their staff to regular information events and workshops related to green finance, or even offer them (see Volz, 2015, p. 42). With these measures the ECB could try to establish a uniform level of knowledge. Banks should become aware of what measures they can implement, and what opportunities and risks those entail. In the second stage with a time-range of 3-4 years, the ECB could introduce a non-compulsory target with regards to the share of green finance in the bank's portfolios (see Volz, 2015, p. 42). Furthermore, the ECB could prompt the banks to select a board member responsible for green finance topics similar to the guidelines of the CBRC. This measure could prevent the issue from drowning among all other challenges that banks experience (see Kirmße, 2017, p. 297-299). Other measures that the ECB could introduce include "(...) annual awards for banks with a high share or a rapidly increasing share of green lending" or even penalty fees, in case of the introduction of compulsory targets (Volz, 2015, p. 42). Nevertheless, these measures contain limits due to the ECB's special position of independence, which will be explained in more detail later.

What the possible measures of the voluntary guidelines all have in common is a communication of a clear definition of sustainability and green finance, avoiding a lack of understanding that "(...) may result in financial markets not realizing their full potential and failing to provide the amount of finance required to reach the EU's long term objectives and global commitments." (Kahlenberg 2017: 1). With the introduction of the Taxonomy Regulation in 2021 progress should be made at this point, since the Taxonomy will make recom-

mendations of for instance green loan criteria (see EU Technical Expert Group on Sustainable Finance, 2020, p. 42).

3.3.2. Disclosure of the ECB carbon impact

Large-scale asset purchases, referred to as Quantitative easing (QE) is a form of unconventional monetary policy. The central bank purchases securities from the open market to increase the money supply and thus encourage lending and investment (see Benford, Berry, Nikolov, & Young, 2009, p. 91-93). The ECB introduced an Asset Purchase Program (APP) in 2014 in line with their unconventional monetary policy. It consists of the corporate sector purchase program (CSPP); public sector purchase program (PSPP); asset-backed securities purchase program (ABSPP) and the third covered bond purchase program (CBPP3) (see European Central Bank, 2020a, p. 1). This abstract will focus on the CSPP. A wide variety of corporate bonds issued by non-bank corporations are eligible for the CSPP. Criteria include the minimum requirement of credit quality 3 on the Eurosystem credit assessment framework (ECAF) and the denomination in euro. Credit quality step 3 is equivalent to an investment grade of BBB- of an external credit rating agency, like for example Standard & Poor's (see European Central Bank, 2020e, p. 1). The securities need to have a remaining maturity of maximal 31 years and minimum 6 months (see De Santis, Geis, Juskaite, Vaz Cruz, et al., 2018, p. 67). There is no minimum bond volume under the CSPP to include the purchase of bonds issued by small companies (see De Santis et al., 2018, p. 67).

In 2017 the research report "The climate impact of quantitative easing" was published by the Grantham Research Institute on Climate Change and the Environment at the London School of Economics (LSE). It states that "(...) low-carbon assets and sectors seem to be relatively underrepresented, with neither the ECB nor the Bank of England purchasing any bonds from renewable energy issuers. "(Matikainen, Campiglio, & Zenghelis, 2017, p. 18).

The study of the LSE also states that there should be a revision of the ECB eligibility criteria, since three of four chosen renewable energy companies are not eligible under the ECB's collateral framework (see Matikainen et al., 2017, p. 14; 20). Issued bonds from corporations belonging to carbon-intensive sectors (e.g. oil, gas, automotive) that are purchased under the CSPP result in a reduction of financing cost for the respective corporations. As a result, their emission of greenhouse gases will not stop and is even further supported through this easy access to funding (see Solana, 2019, p. 556; 562). Jordan and Kalinowski (2019, p. 1) argue that there is misalignment between the policy of the ECB and the climate targets of the European Union. According to them, with the CSPP the ECB reflects the current situation at the corporate bond market "(...) even though financial markets seem misaligned with a mitigation path limiting the global warming to 1,5°." (Jordan & Kalinowski, 2019, p. 1).

The main policy recommendation of the LSE study is to make the whole process more transparent: the ECB should disclose their carbon impact around their purchases (see Matikainen et al., 2017, p. 2; 19). The disclosure could act as a "lead by example" instrument, with positive efforts to the whole financial system and public institutions. With the disclosure of their own carbon impact, the ECB could introduce a voluntary disclosure requirement for banks on their carbon-intensive assets, as a result they might aim for a better capital allocation (see Cecot, 2019, p. 5).

In conclusion, the two introduced measures under the scope of the ECB's soft power show that the ECB can use its convening role to promote the development of sustainable financial market practices (see Volz, 2017, p. 7). By the introduction of voluntary green finance guidelines, the ECB can encourage other financial market participants to address the topic. With a disclosure of their own carbon impact through their CSPP, they would emphasize how important the whole challenge is for them, and underline their own transparency and credibility, which are two characteristics that the ECB considers as crucial (see European Central Bank, 2020k, p. 1).

3.4. The role of the ECB in macroprudential regulation

During the time of the financial crisis in 2008, macroprudential regulation developed to address the risk of the financial system as a whole - referred to as "systemic risk" (see Kenç, 2016, p. 1). The main goal of macro prudential policy is to maintain financial stability.

To achieve this goal, the ECB supervises developments in the banking and financial sectors of the EU together with the other central banks of the EU and the European System of Central Banks (ESCB). Through the application of macro prudential policies, the emergence of possible systemic risks is addressed (see Constâncio, 2016, p. 1-5). Sources for the emergence of systemic risk include asset-price bubbles, excessive risk-taking by banks and excessive debt by companies or households (see European Central Bank, 2017a, p. 1; Constâncio et al., 2019, p. 30). Macro prudential measures of the ECB include different regulations that focus on the capital buffers and the leverage ratio of banks as well as a macro stress testing framework (see Henry et al., 2013, p. 5-9). In the following abstract, the possible role of sustainability-related measures in the macro stress testing framework of the ECB will be discussed.

3.4.1. Climate-related stress tests

Macro stress-testing is a policy tool that is used to identify possible vulnerabilities in the financial system and to measure the degree of systemic risk (see Henry et al., 2013, p. 7-12). It also helps to determine the position of the economy in the financial cycle and eventually the appropriate application of macro prudential policies (see Constâncio, 2016, p. 5-6). Stress tests picture the performance of the banking system in certain scenarios (see Budnik et al., 2019, p. 4). They show if the capital positions of institutions are sufficient and able to preserve financial stability in the system. Hence, they are an important instrument to provide information for regulation authorities (see Budnik et al., 2019, p. 4).

As outlined in chapter 2.2.3, climate-related risks can affect the financial system and financial stability. Therefore, the question arises whether the ECB should include climate-related risks into their stress tests. According to Batten et al. (2016, p. 19), to develop a robust climate-related stress test addressing physical risks, four key factors need to be considered. First, an appropriate stress testing scenario with an impact on the whole financial system needs to be created. Here, the ECB would need to choose a plausible scenario that would affect the whole economy of the EU, including for instance the damage of long-lasting physical damage like infrastructure, large cities with systemic relevance and industrial plants (see Batten et al., 2016, p. 19-20). Second, the most affected sectors of the chosen scenario need to be identified. As highlighted in chapter 2.1, insurance companies may be affected significantly more than other institutions, and since they are of systemic relevance for the financial system, it is important to identify them.

Third, suitable data needs to be gathered to ensure the validity of the test. Fourth, the transmission mechanism of the climate-related shocks across the financial system needs to be illustrated (see Batten et al., 2016, p. 19).

According to Borio, Drehmann, and Tsatsaronis (2012, p. 1), it is difficult to develop stress tests for a specific purpose. As a first step towards a climate-related stress test, the ECB could focus on a specific initial shock related to climate change (see Batten et al., 2016, p. 20-22). One possibility would be to carry out a carbon stress test, based on the scenario of an abrupt repricing of carbon-intensive assets (see Schoenmaker, Van Tilburg, & Wijffels, 2015, p. 24). For the development of a carbon stress test, a sufficient base of information of the carbon-intensity and dependence of sectors is required (see Batten et al., 2016, p. 21). The goal of carbon stress tests would be to determine and quantify the exposure of financial institutions to carbon-intensive assets (see Dikau & Ryan-Collins, 2017, p. 15). Depending on the results of the test, counteract measures like capital buffers could then be imposed to the respective institutions (see Dikau & Ryan-Collins, 2017, p. 15). Crucial for a reliable carbon stress test is a common disclosure standard and measurement procedure, to identify large exposures of carbon intensive assets (see Schoenmaker & Van Tilburg, 2016, p. 2). In 2017, the Financial Stability Board's Task Force on Climate-related Financial Disclosures (TCFD) published recommendations for disclosures related to climate change. They should support companies to understand why the disclosures are important and how they can benefit in terms of their investors' needs (see De Guindos, 2019, p. 1). The Vice President of the ECB argues that "(...) before making it mandatory we need to determine what constitutes high-quality disclosure." (De Guindos, 2019, p. 1).

Moreover, Schoenmaker and Van Tilburg (2016, p. 14) suggest the development of a carbon stress test which is based on the scenario of an abrupt transition towards a low carbon economy. According to them, capital instruments and large exposure restrictions are the most efficient measures for carbon-intensive assets (see Schoenmaker & Van Tilburg, 2016, p. 14). They argue that it is crucial to illustrate scenarios of various sustainability risks, including water shortage, carbon asset bubbles and material shortage, to identify the most critical scenarios "(...) from a financial stability perspective." (Schoenmaker & Van Tilburg, 2016, p. 14).

A few central banks already focused on climate-related stress tests: The Central Bank of Brazil (BCB) requires banks to carry out stress tests that measure the possible impact of environmental catastrophes on their balance sheets in order to check their capital sufficiency (see Dikau & Ryan-Collins, 2017, p. 3-4; 23). The Bank of England (BoE) announced to introduce climate stress tests for the largest banks and insurance companies of the UK in 2021. Key factors that the BoE will consider in their climate-related stress tests are: a longer time horizon since climate risks develop over a long time in comparison to conventional risks, as well as different scenarios that aim to capture an early, a late and no policy introduction (see Bank of England, 2019, p. 1) With the climate stress test, the BoE aims to identify the risk exposure of the institutions rather than to test their capital sufficiency (see Bank of England, 2019, p. 1; 5). The ECB stated that it is currently in the process of developing a framework for conducting climate risk stress tests, which will include physical and transition risks and their interaction (see De Guindos, 2019, p. 1).

3.5. The role of the ECB in micro prudential supervision

In general, macro prudential and micro prudential supervision share some instruments. However, the focus of micro prudential supervision is the stability of individual financial institutions while macro prudential considers the financial stability of the financial system as a unit (see Boissay & Cappiello, 2014, p. 135). While holding ultimate responsibility, the ECB conducts supervisory tasks within the Single Supervisory Mechanism (SSM) (see European Central Bank, 2018, p. 4). The SSM consists of the ECB and the national competent authorities (NCA). It aims to share responsibilities between the ECB and the national competent authorities, while the ECB supervises from a European perspective. The mission statement of the SSM states that "The ECB is responsible for the effective and consistent functioning of the SSM, with a view to carrying out intrusive and effective banking supervision (...)" (European Central Bank, 2018, p. 4). Tasks of the ECB include the development of a common approach of supervision, ensuring the consistent application of regulations and supervisory policy as well as unified supervisory and corrective actions (see European Central Bank, 2020i, p. 1). As a result, the ECB is allowed to "(...) grant or withdraw banking licences, assess banks' acquisition and disposal of qualifying holdings, ensure compliance with EU prudential rules, set higher capital requirements (...)" (European Central Bank, 2020i, p. 1). In general, there is a wide range of financial regulation instruments that the ECB could use to encourage financial institutions to consider climate-related risks and challenges. In the following, disclosure requirements and environmental, social and governance (ESG) standards will be discussed.

3.5.1. Disclosure requirements

As mentioned in chapter 3.4.1, in 2017 the TCFD published recommendations regarding disclosures for financial institutions, highlighting the importance of disclosure for the correct pricing of risks in financial markets (see Task force on climate-related financial disclosures TCFD, 2017, p. 2). They argue that disclosure requirements are a fundamental factor in dealing with climate change and environmental risks. The TCFD emphasizes that a lack of information about risks can result in a mispricing of assets and thus in a misallocation of capital, which eventually can lead to sudden price corrections endangering financial stability (see Task force on climate-related financial disclosures TCFD, 2016, p. 2). As of today, the TCFD encourages companies to apply its recommendations, but there is no mandatory regulation yet. Consequently, the introduction of mandatory disclosure requirements by the ECB for all financial institutions could be a possible solution to make sure that assets are correctly priced. An example for a mandatory disclosure is the article 173 of the Energy Transition Law of France: listed companies need to disclose financial risks related to climate change in their annual report, as well as their measures that aim to reduce those risks and the effects of climate change on their operating activities (see UNEP Finance Initiative , 2016, p. 7). Banks are required to disclose the risks uncovered by stress tests and the risk of excessive leverage (see UNEP Finance Initiative , 2016, p. 7). In addition to that, Volz (2017, p. 14) states that the transparency that comes with disclosure requirements is essential for macro prudential policies, like stress tests, which were discussed in chapter 3.4.1 before.

3.5.2. ESG risk management standards

Environmental, social and governance (ESG) risks refer to the inclusion of non-financial risks which result from "(. . .) environment and sustainability, reputation or brand, legal, technological, product or service quality, labor, ethical conduct, compliance, and strategic considerations." (see Idowu, Capaldi, Zu, & Gupta, 2013, p. 1025). The European Commission aims to put ESG considerations in the centre of the financial system to pursue the goal of transforming the European economy into a circular system (see European Commission, 2018a, p. 1). The role of the ECB at this point could be to require banks and other financial institutions to adopt ESG risk management standards (see Dikau & Volz, 2018, p. 6). Banks and financial institutions could include ESG risk management standards in their governance frameworks to consider possibly damaging environmental effects of financial products and services (see Cecot, 2019, p. 5). Moreover, commercial banks could include ESG risks as a criterium in their credit approval process (see Cecot, 2019, p. 5). As a result, there may be less investments in polluting and carbon intensive companies (see Dikau & Volz, 2018, p. 6).

Recently, the ECB engaged in that topic: in May 2020, the ECB issued a draft guide for banks on climate-related and environmental risk management (see European Central Bank, 2020f, p. 3-7). The guide is not legally binding, but it aims to promote awareness of environmental and climate change related risks in the banking sector. The ECB requires banks to assume responsibility for the risks that can ultimately affect the real economy (see European Central Bank, 2020d, p. 1). In the guide, the ECB formulates its expectations towards banks to include climate-related and environmental risks in their risk management frameworks. Moreover, the ECB highlights that it anticipates banks to engage in environmental disclosures to be more transparent (see European Central Bank, 2020d, p. 1). The ECB asks for comments until September 2020 and is willing to finalize the guide afterwards (see European Central Bank, 2020d, p. 1). Next to its role in micro prudential supervision, the ECB is entrusted with the implementation of monetary policy. In the next abstract, the possible role of the ECB in a more sustainable monetary policy will be discussed.

3.6. The role of the ECB in monetary policy

The overriding priority of the ECB is to maintain price stability (see European Union, 2012, p. 56). The implementation of monetary policy is based on their monetary policy strategy, as outlined in chapter 3.2 (see Verbeken, Rakić, & Patemoster, 2020, p. 2-3). To achieve their goals, the ECB applies different monetary policy instruments. A distinction is made between conventional and unconventional measures. The conventional instruments include open market operations, standing facilities and the holding of minimum reserves (see Verbeken et al., 2020, p. 3-4). To handle the financial crisis and its consequences, the ECB introduced new monetary instruments: unconventional monetary policy instruments (see Potter & Smets, 2019, p. 26). In the following, the main emphasis will be on the unconventional instrument quantitative easing (QE). In 2015, the ECB started their QE program, by buying bonds from commercial banks. The theory states, that the price of bonds increase and thus liquidity increases (see Zappalà, 2018, p. 18). Yields decrease and loans become cheaper, so companies and households can borrow more money to pay back their debt. Hence, investment and spending increase, which ultimately leads to an increase in prices and an inflation rate of close to 2% (see European Central Bank, 2020h, p. 1). In general, the QE is often criticized, but nevertheless the discussion emerged whether QE could aim towards the purchase of green financial assets, like green bonds to support the transition towards a more sustainable economy (see Volz, 2017, p. 16). In the next abstract, a quantitative easing variation focused on the environment, called green quantitative easing (GQE), will be examined.

3.6.1. Green quantitative easing (GQE)

The main procedure of green quantitative easing (GQE) is similarly to general quantitative easing (see Zappalà, 2018, p. 21). The central bank expands its balance sheets by the purchase of assets, hence injecting a great amount of liquidity in the financial system (see Ferron & Morel, 2014, p. 9). The idea of GQE is, that those liquid assets could be used to finance sustainable, environment friendly projects (see Ferron & Morel, 2014, p. 9). The first difference to normal QE

is that central banks are required to purchase inter alia green bonds (see Cecot, 2019, p. 6). A purchase of solely green bonds by the ECB is considered as controversial, since the green bond market is still relatively small (see Sustainable Banking Network (SBN), 2018, p. 11). There is no global legal definition of green bonds yet, but the Green Bond Principles (GBP) provide orientation (see Doran & Tanner, 2019, p. 4). The GBP define Green Bonds as "(...) any type of bond instrument where the proceeds will be exclusively applied to finance or re-finance, in part or in full, new and/or existing eligible Green Projects (...)" (International Capital Market Association (ICMA), 2018, p. 3). Thus, the main concept of the GQE is rather than buying any type of bonds, central banks should purchase bonds that have been issued with the intention to finance projects on sustainable investments (see Dafermos, Nikolaidi, & Galanis, 2018, p. 2). Green bonds could be issued by private as well as public entities, the ECB could buy them on the primary or secondary market, depending on the issuer (see Ferron & Morel, 2014, p. 12). As a result, investors would be stimulated to invest more capital towards sustainable projects (see Dhomps, 2019, p. 32). Moreover, the risk of investing in sustainable projects will decrease since they are often associated with more risk than conventional assets (see Dhomps, 2019, p. 32). The primary objective of GQE would not be economic growth, instead its main purpose would be to support the transition of the economy towards sustainability (see Dafermos et al., 2018, p. 2). Additionally, the GQE program could have a positive impact on job creation, because it can increase economic activity "(...) in the green sector which tends to be more labour intensive." (Dafermos et al., 2018, p. 2, see Dillon 2010: 3-4; see Ferron & Morel, 2014, p. 9). According to a study from the Greenwich Political Economy Research Centre, a GQE program by the ECB could support the fight against climate change: the interest rate on green bonds would decrease in comparison to conventional bonds, as a result investments in sustainable projects would increase (see Dafermos et al., 2018, p. 4). However, the study concludes that many other policies besides to GQE are needed to reach the target of limiting global warming to 2°C as stated in the Paris Agreement (see Dafermos et al., 2018, p. 5). The authors also investigated if a green corporate quantitative easing program could affect climate-related financial stability (see Dafermos et al., 2018, p. 234). Their results demonstrated that "(...) by increasing the price of green corporate bonds, the implementation of such a green QE program can reduce climate-induced financial instability and restrict global warming." (Dafermos et al., 2018, p. 234).

Next to economists like Alain Grandjean, who calls for a the introduction of a GQE program by the ECB, particularly green activists demand for the support of the ECB by financing sustainable and environmental projects (see De Grauwe, 2019, p. 1, see Grandjean, 2016, p. 1). On the opposite, conventional economists argue that the environment is not in the scope of the ECB but within the remit of the government (see De Grauwe, 2019, p. 1). They fear that the ECB will commit itself to injecting too much money in the financial system and therefore fuel inflation (see De Grauwe, 2019, p. 1). This gives a first insight into the ongoing discussions regarding the competences and the mandate of the ECB, that will be outlined in more detail in chapter 4.

3.6.2. Green collateral framework

Commercial banks can borrow money from the central bank under the condition of providing sufficient collateral, examples for eligible collateral are financial securities like bonds (see European Central Bank, 2016, p. 1; see Debelle, 2013, p. 4-7). The collateral aims to protect the balance sheets of central banks against the risk of default by the borrower (see Deutsche Bundesbank, 2018, p. 1). The borrower is the respective commercial bank that obtains liquidity, and if it is unable to pay back the loan, the ECB uses the provided collateral to avert a financial loss (see Blot, Creel, & Hubert, 2018, p. 9). The criteria of eligible collateral are stated in the Eurosystem Collateral Framework (ECF) and the national central banks are responsible to identify eligible assets (see Bindseil, Corsi, Sahel, & Visse, 2017, p. 6, 10). The framework is divided in two frameworks, the "general" and the "temporary" framework, whereas the "temporary framework" overrules the general one (see European Central Bank, 2016, p. 1).

Because collateral is the limit on banks' access to liquidity, it is of particular importance (see Bindseil et al., 2017, p. 3). Commercial banks are granted financing advantages if they pledge sufficient collateral. The assessment of the collateral is based on three criteria to mitigate the risk: "(...) the credit risk associated with the collateral accepted (...) the market risk of an adverse movement in the price of an asset accepted (...) the liquidity risk of an adverse movement in the price of an asset." (European Central Bank, 2015, p. 12). The ECB retains risk discounts on loans to banks. The worse the creditworthiness of the provided collateral, the greater the risk discount. Consequently, a correct valuation of the eligible assets is crucial for an adequate risk management (see Blot et al., 2018, p. 22).

To contribute towards a sustainable financial system, the ECB could revise its collateral eligibility framework. Schoenmaker (2019, p. 7, 16), Jordan et al (2019: 4) and Campiglio et al. (2017, p. 5) propose to integrate low-carbon eligibility criteria into the collateral eligibility framework, to facilitate funding for low-carbon businesses. For commercial banks, environment-friendly assets would become more pleasant, thus the financing conditions for low-carbon companies would improve (see Campiglio et al., 2017, p. 5). Schoenmaker (2019, p. 16) states that the effort to green monetary policy operations could also "(...) give a powerful signalling effect to other financial market participants (...) boosting the case for greening the financial system.". One national bank that already introduced a similar measure is the central bank of Norway (Norges Bank). They have excluded selected companies which are involved in the production of coal-based energy, nuclear weapons and environmental damage from their state pension fund (see Norges Bank, 2020, p. 1).

In terms of their own risk mitigation, the ECB could aim to incorporate climate-change related risks in their collateral framework (see Campiglio et al., 2017, p. 7). As outlined before, climate change poses risks on the financial system and its participants. In order to appropriately measure the default risk of their counter parties, the ECB should therefore take into account climate-related risks into their determination process of eligible assets (see Van Lerven & Bryer, 2020, p. 4). As of today, the assessment by the ECB to evaluate which collateral can be accepted are derived from private sector credit rating agencies, and climate-related risks are not incorporated in their ratings yet (see Van Lerven & Bryer, 2020, p. 4). Consequently, this circumstance grants a relative advantage to carbon-intensive collateral that include climate risks, in comparison to securities that are safer in climate risks terms (see Van Lerven & Bryer, 2020, p. 4). Bolton et al. (2020, p. 54) argue that: "The goal of this proposal is not that central banks should step out of their traditional role when implementing monetary policies, but rather to recognise that the current implementation of market neutrality, because of its implicit bias in favour of carbon-intensive industries (...) could end up affecting central banks' very own mandates in the medium to long term.". Furthermore, Van Lerven and Bryer (2020, p. 5) propose that a greening of monetary policy operations with a long-term approach could support the EU's Green Deal and the transition towards a circular economy by supporting green financial flows. Moreover, the ECB would reduce its own vulnerability to climate-related risks (see Van Lerven & Bryer, 2020, p. 5).

4. Restrictions of the ECB to engage

Despite all the different, presented ways in which the ECB could support the transition towards a sustainable financial system, the special role of the ECB also entails some restrictions. Among them is the role of independence of the ECB, which will be discussed first.

4.1. The role of independence

The independence of central banks refers to "(...) the freedom of monetary policymakers from direct political or governmental influence in the conduct of policy." (see Walsh, 2010, p. 21). As stated in article 282 of the Treaty on the European Union, the ECB "(...) shall be independent in the exercise of its powers and in the management of its finances. Union institutions, bodies, offices and agencies and the governments of the Member States shall respect that independence." (European Union, 2012, p. 121). The political independence of the ECB is essential to pursue its main objective of price stability (see European Central Bank, 2017b, p. 1). Otherwise, politicians would have the incentive of using instruments (e.g. lowering interest rates) of the ECB to influence their political actions which would affect the economy negatively in the long run (see European Central Bank, 2017b, p. 1; see Groepe, 2016, p. 7). To maintain their independence, the ECB needs to ensure their accountability through clear communication and transparency (see Eichengreen et al., 2011, p. 31; see Howarth & Loedel, 2005, p. 123). Economists warn that central banks are putting their independence at risk by taking up the issue of climate change (see Honohan, 2019, p. 2). They argue that central banks were not granted in-dependency to expand their mandate (see Issing, 2019, p. 1). Even if environmental topics may belong to their secondary objectives, central banks should admonish against unrealistic expectations with respect to their contribution towards climate change measures (see Honohan, 2019, p. 2). Groepe (2016, p. 1) argues that the greatest risk to the independence of central banks "(...) is the possible backlash from being unable to deliver on unreasonable expectations.". In concrete terms, they fear that the ECB moves away from its core mission by pursuing political interests and thus loses its credibility and independence, which is essential to achieve their primary objective.

4.2. The connection of climate risk and credit risk

As outlined in chapter 2.2.1, one of the main responsibilities of central banks is to maintain financial stability (see Mishkin, 2005, p. 411). With a sudden change in regulatory, companies that rely on fossil-fuel are exposed to transition risks (see Solana, 2019, p. 572). They may not be able to use fossil fuels anymore (see Bank of England, 2015, p. 51-52). Thus, their profitability and value can be affected negatively (see Bank of England, 2015, p. 50). According to the European Systemic Risk Board (2016, p. 12) fossil-fuel companies are mainly financed by debt. Capasso, Gianfrate, and Spinelli (2020, p.25) argue that companies with a high amount of CO2 emissions are more likely to default. However, many banks have not started yet to take climate risks into account during their credit assessment process (see Colas et al., 2019, p. 16). As a result, an abrupt change can cause credit losses and a repricing of debt, which can ultimately affect financial stability (see Williams & Case, 2016, p. 3-6). Credit risk refers to the risk of a financial loss in case a borrower can not repay their debt to the lender (see Monnin, 2018, p. 3). Consequently, climate risks can be considered a source of credit risk (see Monnin, 2018, p. 2; see Solana, 2019, p. 572). If the ECB now would decide to exclude carbon-intensive companies in its CSPP, or to introduce Green Quantitative Easing (GQE), the cost of financing for carbon-intensive companies would increase (see Solana, 2019, p. 571). Together with possible emerging regulatory burdens like for example a CO2 tax or the general challenge to become sustainable in their operations, the long-term viability of such companies would be endangered (see Solana, 2019, p. 572). Ultimately, the risk of credit loss for commercial banks would emerge since those companies might not be able to repay their debt. This in turn could lead to a risk to financial stability.

4.3. The ECB mandate and conflicting objectives

In the last years, public awareness of climate change increased and a debate on the role of the ECB and central banks

emerged (see Batten et al., 2016, p. 3; see Dhomps, 2019, p. 29). A broad consensus argues that central banks in general can not ignore climate change (see NGFS, 2019b, p. 2, see United Nations Environment Programme, 2016, p. 53). However, there is no exact definition on the extent to which the ECB should consider climate change risks in their frameworks, or what role the ECB should hold in promoting sustainable finance. Researchers started to analyse the ability of the European Union to support environmental objectives as well as the mandate of the ECB with regards to a green monetary policy (see Solana, 2019, p. 575, see Fischer, 2018, p. 1-16). However, the integration of environmental objectives in the ECB's mandate faces criticism. Benoît Cœuré, Member of the Executive Board of the ECB, argues that the second objective of the ECB, to support general economic objectives, can be questioned regarding their orientation. He illustrates the example with the question "(...) why the ECB should not promote industries that promise the strongest employment growth, irrespective of their ecological footprint." (Cœuré, 2018, p. 1). Thereby he outlines the potentially conflicting objectives of the ECB, apart from their main objectives (see Dikau et al. 2019: 7). In general, Cœuré states that the ECB can occupy a promoting and supporting role in decreasing the risks from climate change within their mandate (see Cœuré, 2018, p. 1). On the contrary, the president of the Deutsche Bundesbank, Jens Weidmann, comments critically on possible instruments like GQE (see Weidmann, 2019, p. 3). He points out that the responsibility for a transition towards a low carbon economy can be supported by central banks but remains in the responsibility of politicians (see Weidmann, 2019, p. 4). As a result, the ECB may face restrictions regarding their competences that are in scope of their mandate. For instance, in 2020 the German Federal Constitutional Court (GFCC) judged that the Public Sector Purchase Program (PSPP) of the ECB is not in line with their mandate. They argued that the mandate is strictly limited to conducting monetary policy; and that economic or social policy decisions are not in scope of their mandate, even if they implement monetary policy instruments for this purpose (see Bundesverfassungsgericht, 2020, p. 70). Consequently, one could argue that a similar judgement could take place if the ECB decides to pursue climate policy with monetary policy instruments. Thus, the question arises whether the ECB's mandate should be extended, or if even a separate mandate should be introduced to permit the ECB to pursue climate change and sustainability objectives.

In addition to that, Volz (2017, p. 18) argues that adding environmental objectives to the ECB's tasks, could result in an overburdening. He states that the ECB would have to pursue "(...) to many objectives and have too few tools (...)" and that the ECB would need to introduce productive instruments to pursue environmental goals without affecting other objectives (Volz, 2017, p. 18). One of the ECB's core principles is anchored in Article 127 TFEU: market neutrality (see European Union, 2012, p. 50; 56, see Weidmann, 2019, p. 3). This means that the ECB may not preferentially buy bonds from certain companies or states (see Mersch, 2018, p. 1). In its purchase programs, the ECB needs to purchase the cross-section of the market, to avoid that the program does distort the market (see European Central Bank, 2020c, p. 1). If the ECB would now focus on sustainable "green" bonds, this would no longer be the case. Consequently, a separate mandate for the prosecution of climate would be necessary in terms of legal authorization.

5. An evaluation of instruments – the PCT approach

In the following abstract, a possible approach for the evaluation of the presented instruments will be outlined. The objective of the valuation approach is to evaluate the instruments in a reasonable manner. As an evaluation of the instruments could not be sufficiently quantified, a qualitative evaluation approach is used below to identify the main characteristics. The approach is based on three criteria that will be applied to the respective instrument.

The criterion of proportionality examines the suitability of each instrument to achieve the objective of a more sustainable financial system. For instance, the unconditional pursuit of the objective of sustainability without regard to the main objectives of price stability and financial stability would disregard the principle of proportionality. The criterion is based on the Treaty of the European Union, that states: "(...) Under the principle of proportionality, the content and form of Union action shall not exceed what is necessary to achieve the objectives of the Treaties." (European Union, 2012, p. 6).

The second criterion, competency, investigates whether the ECB acts within its mandate or exceeds its competencies. As stated before, the primary mandate of the ECB is to maintain price stability and an inflation target of below but close to 2%. Without disregard to the objective of price stability, general economic policies in the Union should be supported (see Estella, 2018, p. 78).

The third criterion is the time horizon. Since climate change is a rapidly progressing process, it outlines how quickly the respective instrument could be introduced and could have an impact. A distinction is made here between a short time horizon (up to one year), a medium time horizon (3-5 years) and a long-time horizon (5-10 years). Based on the three criteria, the evaluation model is referred to in the following as the PCT (Proportionality-Competency-Time) approach. In the next abstract, two instruments that were outlined in more detail before will be evaluated with the approach and conclusively assessed.

5.1. An assessment of two instruments using the PCT approach

First, voluntary green finance guidelines and frameworks as described in chapter 3.3.1 will be evaluated. With the aim to promote green finance products like for example green loans, the instrument clearly pursues the objective of a sustainable financial system. Since it is optional and not compulsory, it is expected that no abrupt shifts in the market will occur, so that the objectives of price and financial stability

should not be compromised. Consequently, the criterion of proportionality is fulfilled.

The article 132 of the TFEU allows the ECB to publish recommendations and opinions to fulfil its tasks (see European Union, 2012, p. 58). Besides the main objectives, the ECB's tasks include the support of general economic policies. One could argue that the climate target as stated in the Paris Agreement belongs to those (see Hercelin, 2019, p. 1). Under this assumption, the criterion of competence is also fulfilled. Given its voluntary nature, the introduction of voluntary guidelines should be possible in a short-term horizon as it costs less than binding regulations (see United Nations Environment Programme, 2017, p. 20). With the support of the classifications of the Taxonomy Regulation, the content creation of voluntary guidelines by the ECB should be possible in a short time horizon of approximately one year. All in all, despite the possibility of a quick introduction, it is difficult to estimate what impact voluntary guidelines could have. Further research, for example in the form of surveys to quantify the knowledge and willingness of investors to contribute towards green finance, is needed to assess the success of voluntary guidelines. However, the ECB holds a position as an EU institution to draw the attention of financial market participants to green products. Voluntary guidelines could simplify and accelerate this process.

The next instrument that will be analysed is the unconventional monetary policy instrument Green Quantitative Easing (GQE) as discussed in chapter 3.6.1. By directing the program towards green bonds, the main objective of the instrument would be the transition towards a more sustainable financial system rather than economic growth (see Dafermos et al., 2018, p. 2). However, the market of green bonds is still limited, and a discrimination of non-green bonds could endanger financial stability (see Deschryver & De Mariz, 2019, p. 1-4). Furthermore, the implementation of monetary policy instruments needs to be in accordance with the principle of market neutrality. An asset purchase program focused on green bonds would violate this principle as laid down in article 127 of the TFEU (see European Union, 2012, p. 56). Consequently, the criterium of proportionality is not fulfilled. The German Federal Constitutional Court (GFCC) clearly distinguishes monetary policy from economic policy due to the definition in the TFEU. It rejects actions that pursue economic policy objectives with monetary policy (see Rövekamp, Bälz, & Hilpert, 2015, p. 52-53). According to the GFCC, the responsibility for economic policies is in scope of the member states and not the European Union, the ECB is only allowed to support economic policy objectives, without pursuing its own economic policy (see Bundesverfassungsgericht, 2020, p. 70). Thus, the criterium of competency is not fulfilled.

In March 2020, immediately after the outbreak of the Covid-19 pandemic in Europe, the ECB announced a Pandemic Emergency Purchase Program (PEPP) to counteract risks from the global pandemic (see Deutsche Bundesbank, 2020, p. 1). In addition to that, a quantitative easing program was already implemented by the ECB in 2014, therefore a GQE would be no unfamiliar instrument in general, which could limit the time needed for implementation. Hence a GQE may would be introduced within a short-time horizon of one year, and its duration could vary depending on its measured results. However, the impact of GQE is not significant with regards to the climate target of the Paris Agreement (see Dafermos et al., 2018, p. 4-5). Subsequently one could argue whether the ECB should really introduce such a controversial measure and accept various risks, for relatively small effects.

6. Conclusion

To pursue a sustainable financial system, it will be essential to involve the financial sector, as well as its participants and authorities. This thesis discussed the role the European Central Bank could play in a transition towards a sustainable financial system. It outlined how climate-related risks can spread to the financial system and why central banks are concerned of them. While introducing possible instruments for a proactive role of the ECB in different policy areas, the thesis also pointed out the restrictions and risks of the ECB to engage. Next to the possible endangering of their institutional independence, the mandate of the ECB was investigated. To the extent that the ECB pursues financial stability, including climate change risks into stress-tests and frameworks already belongs to their mandate. However, incorporating climate risks into existing frameworks is a challenging task, since traditional frameworks rely on historical data, and climate change is characterized with deep uncertainty. All in all, it is to question whether macroeconomic models can serve sufficient data and information in this case. Further, detailed research is necessary at this point. However, ignoring the risks from climate change can not be considered an option.

Whether central banks in general should promote green finance is a political issue that is controversial and needs to be examined very carefully. At this point, complex political questions should not be converted into moral certainties. According to TFEU, it is not the task of the ECB to finance green projects or to even implement a GQE: the ECB is not allowed to define its mandate and adapt it to, for example, a European Green Deal. By implementing such controversial measures, the ECB would distort the market. Moreover, a "green" monetary policy could threaten democracy in the EU.

However, the ECB is only one institution among public institutions that can influence the transition towards a sustainable financial system, other relevant institutions are environment ministries, the finance ministry and development banks like the "Credit Institute for Reconstruction" (KfW), who hold an environmental mandate and pursue sustainability objectives. By reviewing financial and macroeconomic instruments and policies, it is important to keep in mind that there are many other types of policies and institutions that could be involved, for instance regulatory emission standards could serve as implicit carbon prices (see Krogstrup & Oman, 2019, p. 33).

Next to the restrictions of the ECB, the thesis presented an evaluation approach for the introduced instruments based

on the criteria of proportionality, competency and time. The aim of the approach was to evaluate the respective instrument in a plausible way bearing in mind the objectives and mandate of the ECB as well as the speed of climate change. Facing these legal and technical challenges, a key role of the ECB to support the transition towards a sustainable financial system could be to provide an appropriate frame of the climate change debate by providing information about the financial risks from climate change. Moreover, the ECB could lead by example by the disclosure of their own carbon impact and by promoting the discussion about climate change as outlined in chapter 3.3. Here, the ECB seems to approach the topic, since May 2020 they are working on the "guide on climate-related and environmental risks", which aims to provide transparency of the ECB's understanding of an appropriate management of climate-related risks (see European Central Bank, 2020g, p. 1-9).

The key points to consider as an answer to the approached research question "What role can the ECB fulfil in a sustainable financial system?" are the following:

- the ECB should not fulfil a proactive role, it should not use controversial instruments to promote a sustainable financial system, it should rather create the framework for the transition by providing information, education and awareness;
- the financing of concepts such as the European Green Deal is within the responsibility of politicians, not the ECB;
- to the extent possible within their mandate, the ECB should manage the risks from climate change on financial stability.

Limiting global warming to below 2°C as determined in the Paris Agreement requires a structural transformation in the economy, and therefore a transition in the areas on which the economy relies. Consequently, the transition will not only require a huge shift towards green, sustainable investments but also shifts in the prices of energy and productive capacity (see Krogstrup & Oman, 2019, p. 39). To conclude the "(...) climate crisis is probably the biggest challenge of our times and it will be faced only if it becomes a priority. Therefore, sustainable finance is only one piece of the green growth puzzle and all economic sectors should now start integrating into their policy environmental considerations and objectives." (Dhomps, 2019, p. 39).

References

Amadeo, K. (2019). *Core Inflation, How It Is Calculated, Its Critical Role.* Retrieved 2020-05-25, from https://www.thebalance.com/core-inflation-rate-3305918

Ansar, A., Caldecott, B., & Tilbury, J. (2013). Stranded assets and the fossil fuel divestment campaign: what does divestment mean for the valuation of fossil fuel assets?

Bank of England. (2015). The impact of climate change on the UK insurance sector. In *Climate Change Adaptation Report.* London: Prudential Regulation Authority.

Bank of England. (2017). *Climate change: what are the risks to financial stability?* Retrieved 2020-05-25, from https://www.bankofengland.co.uk/knowledgebank/climate-change-what-are-the-risks-to-financialstability

Bank of England. (2019). The 2021 Biennial Exploratory Scenario on the Financial Risks from Climate Change. London: Bank of England: Discussion Paper, Financial Policy Committee.

Batten, S., Sowerbutts, R., & Tanaka, M. (2016). Let's talk about the weather: the impact of climate change on central banks. *Staff Working Paper No. 603*.

Benford, J., Berry, S., Nikolov, K., & Young, C. (2009). Quantitative Easing. *Quarterly Bulletin 2009 Q2*.

Bindseil, U., Corsi, M., Sahel, B., & Visse, A. (2017). The Eurosystem collateral framework explained. *Occasional Paper Series No 189*.

Blot, C., Creel, J., & Hubert, P. (2018). The effects and risks of ECB collateral framework changes. *Monetary Dialogue July 2018, Policy Department for Economic, Scientific and Quality of Life Policies*.

Boissay, F., & Cappiello, L. (2014). Micro- versus macro-prudential supervision: potential differences, tensions and complementaries. *Financial Stability Review May 2014*.

Bolton, P., Despres, M., Pereira da Silva, L. A., Samama, F., & Svartzman, R. (2020). The Green Swan. Central banking and financial stability in the age of climate change. *Banque de France*.

Borio, C., Drehmann, M., & Tsatsaronis, K. (2012). Stress-testing macro stress testing: does it live up to expectations? *BIS Working Papers No 369, Monetary and Economic Department*.

Budnik, K. B., Balatti, M., Covi, G., Dimitrov, I., Groß, J., Hansen, I., ... others (2019). Macroprudential stress test of the euro area banking system. *Occasional Paper Series No 226*.

Bundesverfassungsgericht. (2020). *Judgment of the Second Senate of 05 May 2020.* Retrieved 2020-06-17, from https://www.bundesverfassungsgericht.de/SharedDocs/Downloads/EN/2020/05/rs20200505_2bvr085915en.pdf?__blob=publicationFile&v=6

Cambridge Dictionary. (2020). *Taxonomy.* Retrieved 2020-05-25, from https://dictionary.cambridge.org/de/worterbuch/englisch/taxonomy

Campiglio, E., Dafermos, Y., Monnin, P., Ryan-Collins, J., Schotten, G., & Tanaka, M. (2017). Finance and climate change: What role for central banks and financial regulators. In *Submission to Nature Climate Change (Perspective)*. Zurich: Council on Economic Policies.

Cantrell, J. (2020). *Natural Capital.* Retrieved 2020-06-29, from https://www.worldbank.org/en/topic/natural-capital

Capasso, G., Gianfrate, G., & Spinelli, M. (2020). Climate change and credit risk. *EDHEC-Risk Institute Working Paper*.

Carney, M. (2015). Breaking the tragedy of the horizon–climate change and financial stability. *Speech given at Lloyd's of London*.

Carney, M. (2016). Resolving the climate paradox. *Arthur Burns Memorial Lecture*.

CDP. (2019). *Half of Europe's largest firms now link executive pay to climate change.* Retrieved 2020-05-25, from https://www.cdp.net/en/articles/companies/half-of-europes-largest-firms-now-link-executive-pay-to-climate-change

Cecot, M. (2019). Climate Emergency and Central Banks.

Chandavarkar, A. (1987). Promotional role of central banks in developing countries. *IMF Working Paper 87/20*.

China Banking Regulatory Commission. (2012). *Notice of the CBRC on Issuing the Green Credit Guidelines.* Beijing: China Banking Regulatory Commission.

Colas, J., Khaykin, I., Peitsch, A., Schnarr, T., Mettenheimer, P., & Pyanet, A. (2019). *Warum Klimarisiken schon heute für Banken relevant sind.* Retrieved 2020-06-08, from https://www.oliverwyman.de/content/dam/oliver-wyman/v2-de/publications/2019/dec/20191217_Climate_Change_Managing_A_New_Financial_Risk_GERMAN.pdf

Constâncio, V. (2016). Principles of macroprudential policy. *Conference Paper, ECB-IMF Conference on Macroprudential Policy*.

Constâncio, V., Cabral, I., Detken, C., Fell, J., Henry, J., Hiebert, P., ... Salleo, C. (2019). Macroprudential policy at the ECB: Institutional framework, strategy, analytical tools and policies. *Occasional Paper Series*.

Council of the European Union. (2019). *Proposal for a Regulation of the European Parliament and of the Council on the establishment of a framework to facilitate sustainable investment.* Retrieved 2020-05-25, from https://data.consilium.europa.eu/doc/document/ST-14970-2019-ADD-1/en/pdf

Cœuré, B. (2018). *Monetary Policy and Climate Change.* Retrieved 2020-06-09, from https://www.ecb.europa.eu/press/key/date/2018/html/ecb.sp181108.en.html

Dafermos, Y., Nikolaidi, M., & Galanis, G. (2018). Climate change, financial stability and monetary policy. *Ecological Economics*, *152*.

Debelle, G. (2013). *Central bank collateral frameworks and practices.* Basel: Bank of International Settlements.

De Grauwe, P. (2019). *Green money without inflation.* Retrieved 2020-06-04, from https://www.socialeurope.eu/green-money-without-inflation

De Guindos, L. (2019). *Implications of the transition to a low-carbon economy for the euro area financial system.* Retrieved 2020-06-03, from https://www.ecb.europa.eu/press/key/date/2019/html/ecb.sp191121_1~af63c4de7d.en.html

De Santis, R. A., Geis, A., Juskaite, A., Vaz Cruz, L., et al. (2018). The impact of the corporate sector purchase programme on corporate bond markets and the financing of euro area non-financial corporations.

Deschryver, P., & De Mariz, F. (2019). What future for the green bond market? How can policymakers, companies, and investors unlock the potential of the green bond market? *Journal of Risk and Financial Management*.

Deutsche Bundesbank. (2018). *Collateral.* Retrieved 2020-06-20, from https://www.bundesbank.de/en/tasks/monetary-policy/collateral/collateral-625836

Deutsche Bundesbank. (2020). *Pandemic Emergency Purchase Programme (PEPP).* Retrieved 2020-06-17, from https://www.bundesbank.de/de/aufgaben/geldpolitik/geldpolitische-wertpapierankaeufe/pandemic-emergency-purchase-programme-pepp--830356

Dhomps, N. (2019). Advancing green growth in Europe: new approaches for sustainable finance. *Research Paper*.

Diakatos, A., & Bisisidis, S. (2020). *Financial risks stemming from climate change: "Challenging the degree of resilience into a constantly changing environment".* Retrieved 2020-05-25, from https://www2.deloitte.com/gr/en/blog/risk-advisory/2020/financial-risks-stemming-from-climate-change.html

Dikau, S., & Ryan-Collins, J. (2017). Green central banking in emerging market an developing country economies. *London: New Economics Foundation*.

Dikau, S., & Volz, U. (2018). Central banking, climate change and green finance. *ADBI Working Paper 867*.

Doran, M., & Tanner, J. (2019). *Green Bonds – An Overview.* Retrieved 2020-06-04, from https://www.bakermckenzie.com/en/-/media/files/insight/publications/2019/05/green-bonds--an-overview--may-2019.pdf

Eichengreen, B., El-Erian, M., Fraga, A., Ito, T., Pisani-Ferry, J., Prasad, E., ... others (2011). *Rethinking central banking: Committee on International Economic Policy and Reform*. Washington: Brookings Institution.

Estella, A. (2018). *Legal Foundations of EU Economic Governance.* Cambridge: Cambridge University Press.

EU Technical Expert Group on Sustainable Finance. (2020). *Taxonomy: Final report of the Technical Expert Group on Sustainable Finance.* Brussels: European Commission.

European Central Bank. (2009). *The importance of insurance companies for financial stability.* Retrieved 2020-05-27, from https://www.ecb.europa.eu/pub/pdf/fsr/art/ecb.fsrart200912en_05.pdf

European Central Bank. (2011). *The monetary policy of the ECB.* Retrieved 2020-05-27, from https://www.ecb.europa.eu/pub/pdf/other/monetarypolicy2011en.pdf

European Central Bank. (2015). *The financial risk management of the Eurosystem's monetary policy operations.* Retrieved 2020-06-24, from https://www.ecb.europa.eu/pub/pdf/other/financial_risk_management_of_eurosystem_monetary_policy_operations_201507.en.pdf

European Central Bank. (2016). *What is collateral?* Retrieved 2020-06-24, from https://www.ecb.europa.eu/explainers/tell-me/html/collateral.en.html

European Central Bank. (2017a). *A quick guide to macroprudential policies.* Retrieved 2020-06-24, from https://www.ecb.europa.eu/explainers/tell-me-more/html/macroprudentialpolicies.en.html

European Central Bank. (2017b). *Why is the ECB independent?* Retrieved 2020-06-24, from https://www.ecb.europa.eu/explainers/tell-me-more/html/macroprudentialpolicies.en.html

European Central Bank. (2018). *SSM Supervisory Manual European banking supervision: functioning of the SSM and supervisory approach* . Retrieved 2020-06-24, from https://www.ecb.europa.eu/explainers/tell-me-more/html/macroprudentialpolicies.en.html

European Central Bank. (2020a). *Asset purchase programmes.* Retrieved 2020-05-28, from https://www.ecb.europa.eu/mopo/implement/omt/html/index.en.html

European Central Bank. (2020b). *Committee on Economic and Monetary Affairs – Monetary Dialogue with Christine Lagarde, President of the European Central Bank* . Retrieved 2020-07-05, from https://www.bankingsupervision.europa.eu/legalframework/publiccons/pdf/climate-related_risks/ssm.202005_draft_guide_on_climate-related_and_environmental_risks.en.pdf

European Central Bank. (2020c). *Corporate sector purchase programme (CSPP) – Questions & Answers.* Retrieved 2020-06-10, from https://www.ecb.europa.eu/mopo/implement/omt/html/cspp-qa.en.html

European Central Bank. (2020d). *ECB launches public consultation on its guide on climate-related and environmental risks.* Retrieved 2020-06-03, from https://www.bankingsupervision.europa.eu/press/pr/date/2020/html/ssm.pr200520~0795c47d73.en.html

European Central Bank. (2020e). *Eurosystem credit assessment framework (ECAF).* Retrieved 2020-05-28, from https://www.ecb.europa.eu/paym/coll/risk/ecaf/html/index.en.html

European Central Bank. (2020f). *Guide on climate-related and environmental risks.* Retrieved 2020-06-03, from https://www.bankingsupervision.europa.eu/legalframework/publiccons/pdf/climate-related_risks/ssm.202005_draft_guide_on_climate-related_and_environmental_risks.en.pdf

European Central Bank. (2020g). *Guide on climate-related and environmental risks.* Retrieved 2020-07-05, from https://www.bankingsupervision.europa.eu/legalframework/publiccons/pdf/climate-related_risks/ssm.202005_draft_guide_on_climate-related_and_environmental_risks.en.pdf

European Central Bank. (2020h). *How quantitative easing works.* Retrieved 2020-06-04, from https://www.ecb.europa.eu/explainers/show-me/html/app_infographic.en.html

European Central Bank. (2020i). *Single Supervisory Mechanism.* Retrieved 2020-06-03, from https://www.bankingsupervision.europa.eu/about/thessm/html/index.en.html

European Central Bank. (2020j). *Strategy.* Retrieved 2020-05-26, from https://www.ecb.europa.eu/mopo/strategy/html/index.en.html

European Central Bank. (2020k). *Transparency.* Retrieved 2020-05-28, from https://www.ecb.europa.eu/ecb/orga/transparency/html/index.en.html

European Commission. (2017). *Commission guidelines on non-financial reporting.* Retrieved 2020-05-13, from https://ec.europa.eu/info/publications/non-financial-reporting-guidelines_de

European Commission. (2018a). *Proposal for a Regulation of the European Parliament and of the Council on the establishment of a framework to facilitate sustainable investment.* Retrieved 2020-06-03, from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018PC0353&from=EN

European Commission. (2018b). *Technical expert group on sustainable finance (TEG).* Retrieved 2020-05-13, from https://ec.europa.eu/info/publications/sustainable-finance-technical-expert-group_en

European Commission. (2019a). *The European Green Deal.* Retrieved 2020-05-20, from https://ec.europa.eu/info/sites/info/files/european-green-deal-communication_en.pdf

European Commission. (2019b). *Turning challenge into opportunity on the course to becoming the first climate-neutral continent.* Retrieved 2020-05-20, from https://ec.europa.eu/commission/presscorner/detail/en/ac_19_6778

European Commission. (2020a). *Commission proposes a public loan facility to support green investments together with the European Investment Bank.* Retrieved 2020-06-20, from https://ec.europa.eu/commission/presscorner/detail/en/ip_20_930

European Commission. (2020b). *EU Biodiversity Strategy for 2030.* Retrieved 2020-05-23, from https://ec.europa.eu/info/strategy/priorities-2019-2024/european-green-deal/actions-being-taken-eu/eu-biodiversity-strategy-2030_en

European Commission. (2020c). *EU budget for recovery: Questions and answers on the Just Transition Mechanism.* Retrieved 2020-06-20, from https://ec.europa.eu/commission/presscorner/detail/en/qanda_20_931

European Commission. (2020d). *Farm to Fork Strategy – for a fair, healthy and environmentally-friendly food system.* Retrieved 2020-05-23, from https://ec.europa.eu/food/farm2fork_en

European Commission. (2020e). *Questions and Answers on the MFF and Next Generation EU.* Retrieved 2020-06-20, from https://ec.europa.eu/commission/presscorner/detail/en/QANDA_20_935

European Council. (2020). *Sustainable finance: Council adopts a unified EU classification system.* Retrieved 2020-05-27, from https://www.consilium.europa.eu/en/press/press-releases/2020/04/15/sustainable-finance-council-adopts-a-unified-eu-classification-system/

European Parliament. (2020). *Die Politik der EZB wirkt, aber neue Herausforderungen erfordern neue Antworten.* Retrieved 2020-05-26, from https://www.europarl.europa.eu/news/en/press-room/20200206IPR72014/ecb-policy-is-working-but-new-challenges-need-new-responses

European Systemic Risk Board. (2016). Too late, too sudden: Transition to a low-carbon economy and systemic risk. *Reports of the Advisory Scientific Committee*(6 February 2016).

European Union. (2012). *Consolidated version of the Treaty on the Functioning of the European Union.* Retrieved 2020-05-25, from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:12012E/TXT&from=EN

Eurostat. (2019). *HICP Methodology, Statistics explained.* Brussels: European Commission.

Ferrer, R. (2019). Climate change, the green transition and the financial sector. *Dossier Stopping Climate Change: Now or never*.

Ferron, C., & Morel, R. (2014). Smart unconventional monetary (SUMO) policies: giving impetus to green investment. *Climate Report No. 46*.

Fischer, Y. (2018). Global Warming: Does the ECB Mandate Legally Authorise a "Green Monetary Policy?". In *Scaling up Green Finance: The Role of Central Banks.* Berlin: Council on Economic Policies.

Fisher, P., & Alexander, K. (2019). Climate change: the role for central banks. *Working paper No. 2019/6, Data Analytics for Finance & Macro Research Centre*.

Gerdesmeier, D. (2009). Price stability: Why is it important for you?

Giuzio, M., Krušec, D., Levels, A., Melo, A. S., Mikkonen, K., Radulova, P., et al. (2019). Climate change and financial stability. *Financial Stability Review May 2019*.

Goodhart, C. (1995). *The central bank and the financial system*. London: Palgrave Macmillan.

Goodman, D., & McCormick, L. (2019). *Quantitative Tightening to End as Central Banks Retreat.* Retrieved 2020-05-27, from https://www.bloomberg.com/news/articles/2019-07-23/quantitative-tightening-to-end-as-central-banks-sound-retreat

Grandjean, A. (2016). *A Green QE to rebuild Europe: Redirecting ECB's Quantitative Easing towards a Green Transition.* Retrieved 2020-06-04, from https://alaingrandjean.fr/2016/03/14/green-qe-rebuild-europe-quantitative-easing-towards-green-transition/?lang=en

Groepe, F. (2016). The changing role of central banks. *BIS central bankers' speeches*.

Gründl, H., Gal, J., & Dong, M. (2016). The evolution of insurer portfolio investment strategies for long-term investing. *OECD Journal: Financial Market Trends*, *2016*(1).

Henry, J., Kok, C., Amzallag, A., Baudino, P., Cabral, I., Grodzicki, M., ... others (2013). A macro stress testing framework for assessing systemic risks in the banking sector. *Occasional Paper Series No 152*.

Hercelin, N. (2019). *Green monetary policy: it's about changing the mindset, not the mandate.* Retrieved 2020-06-17, from https://www.positivemoney.eu/2019/04/green-ecb-change-mindset-not-mandate/

Honohan, P. (2019). Should Monetary Policy Take Inequality and Climate Change into Account? *Working Paper 19-18*.

Howard, A., & Hassler, M. (2018). *How will physical risks of climate change affect companies?* Retrieved 2020-05-25, from https://www.schroders.com/fr/insights/economics/how-will-physical-risks-of-climate-change-affect-companies/

Howarth, D., & Loedel, P. (2005). *The European Central Bank: The New European Leviathan?* (2nd ed.). New York: Palgrave Macmillan.

Idowu, S. O., Capaldi, N., Zu, L., & Gupta, A. D. (2013). *Encyclopedia of corporate social responsibility*. Berlin: Springer Verlag.

International Capital Market Association (ICMA). (2018). *Green Bond Principles.* Retrieved 2020-06-04, from https://www.icmagroup.org/green-social-and-sustainability-bonds/green-bond-principles-gbp/

International Monetary Fund. (2004). *Compilation Guide on Financial Soundness Indicators.* Retrieved 2020-05-27, from https://www.imf.org/external/np/sta/fsi/eng/2004/guide/chap2.pdf

IPCC. (2018). Zusammenfassung für politische Entscheidungsträger. *1,5 °C globale Erwärmung. Ein IPCC-Sonderbericht über die Folgen einer globalen Erwärmung um 1,5 °C gegenüber vorindustriellem Niveau und die damit verbundenen globalen Treibhausgasemissionspfade im Zusammenhang mit einer Stärkung der weltweiten Reaktion auf die Bedrohung durch den Klimawandel, nachhaltiger Entwicklung und Anstrengungen zur Beseitigung von Armut*.

Issing, O. (2019). *Central banks aren't responsible for fighting climate change – governments are.* Retrieved 2020-06-05, from https://www.weforum.org/agenda/2019/12/this-is-the-problem-with-central-banks-fighting-climate-change

Ivlevia, D. (2017). *Stranded Assets: Wirtschaftliche Risiken fossiler Rohstoffe und Handlungsoptionen für die internationale Zusammenarbeit.* Retrieved 2020-05-27, from https://www.adelphi.de/de/projekt/stranded-assets-wirtschaftliche-risiken-fossiler-rohstoffe-und-handlungsoptionen-f%C3%BCr-die

Jordan, S., & Kalinowski, W. (2019). Aligning monetary policy with the EU's climate targets. *Veblen Institute for Economic Reforms*.

Kenç, T. (2016). Macroprudential regulation: history, theory and policy. *BIS Papers No 86*.

Kirmße, S. (2017). Die Entwicklungen europäischer Banken im Spannungsfeld von Niedrigzins, Regulatorik und Digitalisierung. Wiesbaden: Springer Gabler.

Krogstrup, S., & Oman, W. (2019). Macroeconomic and financial policies for climate change mitigation: A review of the literature. *IMF Working Paper, Monetary and Capital Markets Department*.

Lagarde, C. (2020). *Hearing at the Committee on Economic and Monetary Affairs of the European Parliament.* Retrieved 2020-05-26, from https://www.ecb.europa.eu/press/key/date/2020/html/ecb.sp200206~edb83d06a3.en.html

Loan Market Association. (2018). *Green Loan Principles.* Retrieved 2020-05-27, from https://www.lma.eu.com/application/files/9115/4452/5458/741_LM_Green_Loan_Principles_Booklet_V8.pdf

Matikainen, S., Campiglio, E., & Zenghelis, D. (2017). The climate impact of quantitative easing. *Policy Paper*.

McKibbin, W. J., Morris, A. C., Panton, A., & Wilcoxen, P. (2017). Climate change and monetary policy: Dealing with disruption. *Climate and Energy Economics Discussion Paper*.

Mersch, Y. (2018). *Climate change and central banking.* Retrieved 2020-05-28, from https://www.ecb.europa.eu/press/key/date/2018/html/ecb.sp181127.en.html

Mishkin, F. S. (2005). *The Economics of Money, Banking and Financial Markets*. Toronto: Pearson.

Monnin, P. (2018). Integrating climate risks into credit risk assessment-current methodologies and the case of central banks corporate bond purchases.

NGFS. (2019a). *A call for action. Climate change as a source of financial risk.* Retrieved 2020-05-26, from https://www.ngfs.net/sites/default/files/medias/documents/synthese_ngfs-2019_-_17042019_0.pdf

NGFS. (2019b). *A sustainable and responsible investment guide for central banks' portfolio management.* Retrieved 2020-05-26, from https://www.ngfs.net/sites/default/files/medias/documents/ngfs-a-sustainable-and-responsible-investment-guide.pdf

Norges Bank. (2020). *Observation and exclusion of companies.* Retrieved 2020-06-24, from https://www.nbim.no/en/the-fund/responsible-investment/exclusion-of-companies/

Nye, J. (2017). Soft power: the origins and political progress of a concept.

Parker, M. (2016). The impact of disasters on inflation. *Working Paper Series No 1982*.

Parker, M. (2017). Global inflation: the role of food, housing and energy prices. *Working Paper Series No 2024*.

Potter, S., & Smets, F. (2019). Unconventional monetary policy tools: a cross-country analysis. *CGFS Papers No 63*.

Qingrong, W., Wang, L., Yiting, S., & Lina, G. (2013). The Sustainable Performance of Chinese Banking and International Financial Institutions. *WWF Report*.

Rövekamp, F., Bälz, M., & Hilpert, H. G. (2015). Central banking and financial stability in East Asia. *Financial and Monetary Policy Studies*, *40*.

Savage, K. (2019). *Rhode Island: Climate Suit Vs. Big Oil Does Not Involve Federal Claims.* Retrieved 2020-05-26, from https://www.climateliabilitynews.org/2019/12/27/rhode-island-climate-suit-appeal/

Schnabel, I. (2020). *Interview with die welt.* Retrieved 2020-05-26, from https://www.ecb.europa.eu/press/inter/date/2020/html/ecb.in200211~d2b4afb5d3.en.html (26th May 2020)

Schoenmaker, D. (2019). Greening Monetary Policy. *Working Paper Issue 02*.

Schoenmaker, D., & Van Tilburg, R. (2016). What role for financial supervisors in addressing environmental risks? *Comparative Economic Studies*.

Schoenmaker, D., Van Tilburg, R., & Wijffels, H. (2015). What role for financial supervisors in addressing systemic environmental risks. *Sustainable Finance Lab Working Paper*.

Solana, J. (2019). The Power of the Eurosystem to promote environmental protection. *European Business Law Review*.

Stolze, M., Sanders, J., Kasperczyk, N., Madsen, G., & Meredith, S. (2016). CAP 2014-2020: Organic farming and the prospects for stimulating public goods.

Sustainable Banking Network (SBN). (2018). *Creating Green Bond Markets – Insights, Innovations, and Tools from Emerging Markets.* Retrieved 2020-06-04, from https://www.climatebonds.net/files/reports/sbn-ifc-cbi_creating-green-bond-markets_report-2018-.pdf

Task force on climate-related financial disclosures TCFD. (2016). *Recommendations of the Task Force on Climate-related Financial Disclosures.* Retrieved 2020-06-07, from https://www.fsb-tcfd.org/

wp-content/uploads/2016/12/16_1221_TCFD_Report_Letter.pdf

Task force on climate-related financial disclosures TCFD. (2017). *Recommendations of the Task Force on Climate-related Financial Disclosures.* Retrieved 2020-06-07, from https://www.fsb-tcfd.org/wp-content/uploads/2017/06/FINAL-2017-TCFD-Report-11052018.pdf (7th June 2020)

Thiele, A. (2013). *Das Mandat der EZB und die Krise des Euro: eine Untersuchung der von der EZB im Zusammenhang mit der Eurokrise ergriffenen Maßnahmen auf ihre Vereinbarkeit mit den rechtlichen Vorgaben des europäischen Primärrechts.* Tübingen: Mohr Sierbeck Verlag.

Thwaites, J., Whitley, S., Wright, H., & Ott, C. (2018). Aligning Finance Is the Forgotten Goal of the Paris Agreement, But It Is Vital to Successful Climate Action.

UNEP Finance Initiative . (2016). *French energy transition law.* Retrieved 2020-06-20, from https://www.unepfi.org/fileadmin/documents/PRI-FrenchEnergyTransitionLaw.pdf (20th June 2020)

United Nations. (2020). *Flagship UN study shows accelerating climate change on land, sea and in the atmosphere.* Retrieved 2020-05-25, from https://news.un.org/en/story/2020/03/1059061

United Nations Environment Programme. (2016). *The Financial System We Need: Aligning the Financial System with Sustainable Development*. Geneva: International Environment House.

United Nations Environment Programme. (2017). *Green finance progress report.* Retrieved 2020-05-27, from http://unepinquiry.org/wp-content/uploads/2017/07/Green_Finance_Progress_Report_2017.pdf

United Nations Framework Convention on Climate Change . (2020). *The paris agreement.* Retrieved 2020-05-25, from https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement (25th May 2020)

Van Lerven, F., & Bryer, N. (2020). The ECB and climate change: outlining a vision for success. *Policy Briefing*.

Verbeken, D., Rakić, D., & Patemoster, D. (2020). *European Monetary Policy.* Retrieved 2020-06-04, from https://www.europarl.europa.eu/ftu/pdf/en/FTU_2.6.3.pdf (4th June 2020)

Verkuijl, C., Piggot, G., Lazarus, M., van Asselt, H., & Erickson, P. (2018). Aligning fossil fuel production with the Paris Agreement.

Volz, U. (2015). Towards a sustainable financial system in Indonesia.

Volz, U. (2017). On the role of central banks in enhancing green finance.

Von Voss-Wittig, H. (2006). *Deutscher Bundestag – Wissenschaftliche Dienste: Soft Power.* Retrieved 2020-05-27, from https://www.bundestag.de/resource/blob/189706/8c40cb75069889f8829a5a0db838da1f/soft_power-data.pdf

Walsh, C. E. (2010). *Central Bank Independence.* London: Palgrave Macmillan.

Weidmann, J. (2019). Climate change and central banks. *Speech at Deutsche Bundesbank's second financial market conference*.

Williams, J., & Case, P. (2016). *Climate policy risk: How does this become a credit risk for banks?* Retrieved 2020-06-08, from https://www.pwccn.com/en/migration/pdf/climate-policy-dec2016.pdf

Zappalà, G. (2018). *Central Banks' role in Responding to Climate Change: Monetary Policy and Macroprudential Regulation* (Unpublished doctoral dissertation). Padua: University of Padova.

Junior Management Science 6(3) (2021) 489-506

Junior Management Science

journal homepage: www.jums.academy

Corporate Taxation in the U.S. and Canada – A Comparative Analysis

Elif Gürlek

Ludwig-Maximilians-Universität München

Abstract

As the degree of global trade integration increased, corporate taxation became a border-crossing matter that obliged governments to reassess the tax attractiveness of their jurisdictions. The U.S. and Canada are two major players of today's trade landscape and the corporate tax environment of these two countries impact MNEs investment decisions. Historically, Canada offered a more favourable tax environment compared to its neighbour. However, the U.S. tax reform, TCJA, challenged Canada's tax attractiveness. This paper aims to assess the similarities and differences of both countries' tax systems after the tax reform based on the tax attractiveness criteria. Following, the paper examines lessons that can be derived for Canada to regain its strong position in the global tax attractiveness scenery. The U.S. and Canada have the potential to set an example for lawmakers and show that it is possible to create a corporate taxation environment that preserves governments' interest whilst creating attractive taxation policies in the eye of MNEs.

Keywords: Corporate taxation; United States; Canada; tax attractiveness; Tax Cuts and Jobs Act.

1. Introduction

"Taxes are what we pay for a civilised society".[1] Taxes are the main source of revenue for governments to create a state of social welfare for their citizens. Despite this, taxes still do not have positive connotations in both individual and corporate taxpayers' minds.

In our increasingly globalised world, trade hardly remains within one country's borders. To minimise their tax liabilities and to exploit the deficiencies in tax systems, multinational enterprises (MNE) are dispersing their business operations around the globe. Since tax payments and related expenditures on corporate income account for a great percentage of a company's expenses, taxation is a central aspect of investment decisions. Thus, countries need to adjust their fiscal policies to increase their tax attractiveness.

The main motivation of this thesis is to understand which country offers a more favourable fiscal environment for MNEs by looking at the dynamics behind international tax competition between countries. Alongside being two dominant players of the global trade landscape, the U.S. and Canada are ideal candidates for a cross-national comparison due to their geographical proximity and close economic links. Since they are both federal states, these countries are also very similar on a fundamental level with regard to their political systems. This similarity continues throughout taxation policies. However, over the years differences in aspects like headline tax rates allowed Canada to gain significant corporate tax attractiveness advantage over the U.S., and to create a more favourable investment environment. This advantage partly faded in 2018 when the Tax Cuts and Jobs Act (TCJA) came into force in the U.S., which was a major tax reform that drew tremendous public attention. TCJA has not only changed the decades-old taxation policies of the U.S. but also had significant implications on an international level. Considering that Canada is one of the countries that is on the frontline when dealing with TCJA, it is almost forced to respond to the reform of its neighbour.

Academia, businesses and civil society have been discussing the U.S. tax reform and its implications for a long time. Various research papers and articles were written in academic journals in recent years. The prevailing argument is that the corporate tax advantage that Canada has enjoyed for decades is in jeopardy and significant implications are on the horizon. Thus, Canada should respond to the U.S. tax

[1] This statement was delivered by Oliver Wendell Holmes Jr. in a United States Supreme Court opinion in 1927 (Compania General de Tabacos de Filipinas v. Collector of Internal Revenue).

DOI: https://doi.org/10.5282/jums/v6i3pp489-506

reform in a grounded and consistent manner. In this regard, the similarities between two tax systems give Canadian policymakers an opportunity to learn from the pivotal elements of the U.S. tax reform.[2]

The principal contribution of this bachelor thesis is the comparison of the tax systems of the U.S. and Canada with regard to corporate taxation from the perspective of their tax attractiveness. Since, as yet, the corporate taxation of both countries has not been analysed by focusing on the similar and diverging aspects, this thesis aims to close the existing gap in the literature. Thus, the question that is being asked is: What are the similarities and differences of MNEs' income taxation in the U.S. and Canada after the U.S. tax reform? Considering how similar corporate taxation in the U.S. and Canada is, it is also crucial to ask: When moving forward, what can Canada learn from the U.S.?

The comparison criteria to assess the similarities and differences are defined on the basis of three indexes: Tax Attractiveness, Tax Complexity and Tax Competitiveness Index. More details with regard to the indexes are going to be discussed in the following section when describing the prior literature written on the topic. These three indexes have a very large number of parameters, whereby many of them are common. For the purposes of this thesis, 12 criteria are chosen in total. Firstly, the parameters are filtered based on the fact of whether they are related to corporate taxation or not. Following, among those filtered parameters, the ones that came across in all three indexes are chosen.[3] In order to create a clearer overview, the parameters are grouped into four categories, whereby each category has been further divided into two to five subcategories. Firstly, "basic elements" are separated into corporate income tax rate and group taxation regime. The second category is "cost recovery" which has been subcategorised as capital cost recovery, loss carryforward and loss carryback. The category "tax incentives" is separated into research and development and patent box regime. Lastly, "international tax regulations" covers general anti-avoidance rules (GAAR), controlled foreign corporation (CFC) rules, thin capitalisation rules, transfer pricing rules and treaty network.

The structure proceeds as follows: Following this introduction, this thesis starts by giving an overview of the taxation system of both countries and by briefly describing the prior literature. Parts of a tax system can only be accurately assessed in the context of the whole system. Therefore, it is important to understand how both countries' legal systems are structured before examining any discrete aspects. In the third part of the thesis, the criteria of corporate taxation are going to be compared.[4] The last part answers the second part of the research question and derives lessons for Canada on the basis of comparison criteria from the previous section. The thesis concludes by summarising the findings of the previous two sections.

2. Overview and Current Developments

2.1. Prior Literature

The "Canada – U.S. Tax Comparisons" by Shoven and Whalley[5] is the only existing comprehensive comparison study so far. It discusses the similarities and differences of taxation systems of both countries beyond the taxation of corporate income. The book consists of several essays, whereby it does not define any criteria to compare corporate taxation and to assess the tax attractiveness of both countries. Even though the book is a good start to understand the fundamental concepts, it has lost its relevance in some respects due to ever-changing tax systems of the U.S. and Canada.

In addition, well-known journals such as the Canadian Tax Journal have published articles that compare certain corporate taxation aspects of both countries. In general, however, the number of comparison studies as such is scarce, and they do not provide a comprehensive comparison of corporate taxation systems.[6] Instead of comparing corporate taxation to assess the tax attractiveness, it has been a more popular research topic to look at the international spillover effects of TCJA, specifically its effects on Canada.[7]

As described in the introduction, the criteria derived from the three indexes lay the groundwork of this thesis. Tax Attractiveness Index aims to make an overall assessment of a country's tax environment to evaluate how attractive it is for corporate investment decisions.[8] Tax Complexity Index covers the criteria that signal the complexity of the tax code and tax framework which are important investment decision factors for MNEs.[9] Lastly, Tax Competitiveness Index measures the competitiveness and neutrality of countries' tax codes among the Organisation for Economic Co-operation and Development (OECD) countries.[10] The parameter of these in-

[2] See McKenzie and Smart (2019b, pp. 3–4).

[3] See the Appendix 1 for a detailed explanation of the methodology followed for the comparison.

[4] See the Appendix 2 for a comprehensive overview of the results of the comparison.

[5] Shoven and Whalley (1992).

[6] Clark and Klemm (2015) highlights how important corporate income tax rate for cross-national investments is, and looks at the rates of major economies, including the U.S. and Canada. Farrar "Canada's Thin Capitalization Rules: The Need for a Subjective Approach" (2006) compares the thin capitalisation rules of Canada with the G-7 countries, including the U.S.McKenzie/Smart "Policy Forum: Business Tax Reform in the United States and Canada" (2019) focuses on the most significant aspect of corporate taxation, the tax rate, and compares it before and after TCJA. Zodrow "Policy Forum: Corporate Income Taxation in Canada" (2008) examines again the corporate income taxation policy of Canada and compares it to the U.S. with regard to its competitiveness.

[7] Bazel, Mintz, and Thompson (2018) emphasises that the new tax reform of the U.S. introduced many ground-breaking changes that will make the country a much more attractive jurisdiction to invest in. Thus, the authors suggest a review of the Canadian tax system. Harris, Keen, and Li (2019) discusses provisions of TCJA that will affect Canada. Alongside fundamental and conventional components of the tax reform, the effects of newly introduced provisions such as the foreign-derived intangible income and the global intangible low-taxed income are considered. Lastly, the article also gives suggestions for possible responses.

[8] See Schanz, Keller, Dinkel, Fritz, and Grosselfinger (2017, pp. 3–4).

[9] See Hoppe, Schanz, Sturm, and Sureth-Sloane (n.d.).

[10] See Bunn and Asen (2019, pp. 1–2).

dexes would allow an objective and comprehensive comparison of countries' tax systems from the standpoint of their tax attractiveness. Nevertheless, a study as such has not yet been conducted, hence this thesis will be the first in this area. Considering the importance of an attractive tax environment as a decisive factor for investments, this thesis will show the stance of the U.S. and Canada and it would also help MNEs that are planning on expanding their business operations in North America to make wiser investment decisions. Additionally, the research would be a good starting point for policymakers who are trying to improve the tax attractiveness of their countries.

2.2. Taxation in the U.S.

The Constitution-based federal republic of the U.S. gives all levels of government, federal, state and municipality, the authority to legislate and regulate. The taxation of income occurs in the U.S. separately on federal and state level, whereas taxing power of states is not absolute and depends on the restrictions of the Congress.[11] The taxes paid on the state and local level are fully deductible when computing federal income tax to avoid duplications.[12]

Tax is by definition the transfer of a portion of the income to the government and it does not give the taxpayer the right to a *quid pro quo* in return.[13] Among the several elements of the U.S. federal tax system, income taxes are the primary component. The tax liability is calculated on the taxable income. However, the U.S. legal system does not provide a single unified definition of income and is intended to "capture all income without committing itself to a specific definition of what it means".[14] This is partly the result of individually added new provisions to the Internal Revenue Code (IRC) to achieve a specific policy role without taking their interaction with the existing provisions into consideration. Due to the fact that overlapping and duplicated provisions continue throughout the overall federal tax system, the taxation system of the U.S. is viewed as complex and multidimensional.[15]

Public Law No. 115-97, publicly known as the Tax Cuts and Jobs Act, was signed by President Donald Trump in 2017. TCJA is the most comprehensive tax reform of the U.S. in over three decades. The new or changed regulations affected every taxpayer in the country, whereby TCJA's most prominent and fundamental changes relate to corporate income taxation rules and the tax treatment of pass-through corporate income. In the near future, it is expected that the law will stimulate the economy positively, however it remains questionable whether the long-term impact will be as positive as lawmakers hoped.[16]

Despite the existence of merely two fundamental tax regimes, worldwide and territorial taxation, corporate taxation differs to a great degree between countries and offers little uniformity. Until the enactment of TCJA in 2018, the U.S. was one of the few remaining developed countries who still implemented worldwide taxation. The pure worldwide taxation regime taxes all income earned by corporations domiciled in the home country. However, the U.S. regime included alterations from the pure form and allowed MNEs to defer the taxation of the foreign-earned income until this income has been repatriated and made available to the parent company. Despite the benefits granted by the deferred taxation rules, the far higher corporate income tax rate of 35 per cent compared to most of the other OECD countries put the U.S. in a disadvantageous position. This also created an unfavourable investment environment for MNEs in the U.S. and incentivised them to engage in aggressive tax planning strategies.[17]

The cession of the worldwide taxation system was one of the most significant changes that was embarked with TCJA. Notwithstanding, lawmakers opted for a modified territorial regime instead of the pure territorial form since the latter would have created strong incentives to shift the taxable income and real investment activity to low-tax jurisdictions and to shift deductions to the U.S. Currently, the foreign-source income is free from taxation in the U.S. and is only taxed in the jurisdiction where the revenue is derived from. Since the repatriation of the foreign earnings does not create a tax liability in the U.S. any longer, under the new territorial regime corporations are more inclined to repatriate their earnings.[18] The changes in MNEs' tax structuring have been seen immediately after TCJA, for instance, Apple announced in the first quarter of 2018 that it would repatriate 94 per cent of its foreign income.[19]

Prior to TCJA, Internal Revenue Service tried to limit profit shifting by regulating transfer pricing rules. Nevertheless, their efforts were not successful, since among 40 per cent of globally profit shifting, U.S. based MNEs shift comparatively more than MNEs from other countries.[20] Thus, the Base Erosion and Anti-Abuse Tax was introduced, which is a provision that applies only to large MNEs and imposes minimum tax on otherwise tax-deductible payments between a U.S. corporation and its subsidiary. In addition, TCJA also establishes other provisions that aim to limit destructive profit shifting such as the foreign-derived intangible income (FDII) and global intangible low-taxed income (GILTI).

Another significant change that was introduced with TCJA is for pass-through businesses. Since the income of the latter is taxed as the owner's personal income at the individual rate[21], their income is not subject to the corporate

[11] See Brackman (2016, pp. 1–3.).
[12] See Pomerleau (2018).
[13] See Burman and Slemrod (2013, p. 6); Sherlock and Marples (2019, p. 5).
[14] McCaffery (2012, p. 36).
[15] See Brackman (2016, pp. 165–166).
[16] See Gale, Gelfond, Krupkin, Mazur, and Toder (2018, p. 590).
[17] See Fischer (2015, pp. 1028–57).
[18] See Yang, Lord, and Saito (2019, pp. 69–83).
[19] See Wakabayashi and Chen (2018) as cited in Yang et al. (2019, pp. 74–75).
[20] See Tørsløv, Wier, and Zucman (2018, pp. 3–4); Tax Policy Center (n.d.).
[21] See Merriam-Webster (n.d.).

income tax rate. Though TCJA changes apply to all corporations including pass-through entities, specific modifications in form of deductions are introduced which will expire after 2025.

2.3. Taxation in Canada

Income tax was first introduced as a temporary measure in Canada to raise revenue for the World War I. However, it continued after the war and income tax has become the single most important revenue-raising method in the country. In the meantime, the income taxation proved to be useful since it became a mechanism for centralising power in the federal government which has been previously reserved to the provinces. Over the years, Canadian lawmakers continued to refine their Income Tax Act (ITA) to cope with the challenges brought by globalisation for its corporate taxpayers.[22]

Canadian income tax system is composed of a multitude of federal, provincial and municipal taxes. Taxing power is divided between the federal government and its political subdivisions based partially on the differentiation between direct and indirect taxation, whereby the federal government is responsible for direct taxation and there is no constitutional limit to the federal taxing power. Despite this, there is no attempt to make a distinction between direct and indirect taxation in the statute. As a result, if the legislation is firmly decided on describing a tax as “direct", there is the possibility of declaring any tax as direct.[23] Provinces have then the power to levy the direct taxation on the income earned in their jurisdiction, whereby all of the 10 provinces and the three territories levy income tax. The federal government of Canada collects the tax on behalf of the provincial governments according to the tax collection agreement between them. Even though provincial governments must adopt the federal income tax base in consideration, they apply their own tax base and provide tax credits.[24]

3. Comparison of Corporate Taxation in the U.S. and Canada

3.1. Basic Elements

3.1.1. Corporate Income Tax Rate

Corporate income tax is the direct tax on the profits of a corporation. It describes the rate at which each dollar of profit is taxed. The enforcement of high corporate income taxes by the governments reduces the after-tax rate of return of corporate investments and increases the cost of capital. As a result, companies become reluctant to invest and the overall economic output decreases. In a broader sense, a greater tax burden also leads to higher prices for end-consumers and reduces the salaries of employees. Therefore, the rate at which the corporate income is taxed has a remarkably significant impact on a country’s economy, and governments that offer lower corporate income tax rates are seen as more attractive.[25]

The taxation of corporate income occurs in the U.S. both on the federal and state level. Prior to the enactment of TCJA, the federal corporate income tax rate was 35 per cent and was notably high compared to other OECD countries. In order to increase the competitiveness of U.S. based MNEs and re-establish the tax attractiveness of the country, the federal corporate income tax rate was reduced to 21 per cent. Also, progressive taxation on the federal level was eliminated, thus the federal corporate income tax rate is a flat rate for all companies regardless of their income.[26]

In accordance with the similarity of their governmental organisation, both the U.S. and Canada impose taxes on both the federal levels and on the level of their political subdivisions. In Canada, these subdivisions are provinces and territories. Canadian federal corporate income tax rate is 15 per cent after the general tax deduction, which is considerably lower than the headline rate in the U.S. General tax deduction is applicable for corporations that are not subject to preferential treatment. For instance, Canadian-controlled private corporations are not allowed to utilise this deduction.

Different from the U.S., Canada imposes progressive taxation on corporate income, whereas the first 500.000,00 CAD (600.000,00 CAD in Saskatchewan) of active business income is subject to a net federal rate of 9 per cent.[27] Progressive taxation is seen as a mechanism for governments to support their small businesses and is an attractive feature of a country’s tax landscape. Thus, Canada has an advantage on this matter compared to the U.S. The corporate income taxes imposed by political subdivisions of both countries vary greatly, whereby the combined corporate income tax rate is on average 26,47 per cent in Canada and 25,77 per cent in the U.S.[28]

3.1.2. Group Taxation Regime

Generally, under income tax law, a company is treated as a single separate taxable unit. This policy is derived from the traditional separate entity doctrine, under which a company is examined as a distinct legal entity and separate from its shareholders. However, with globalisation, the landscape of the trade has changed, and corporate groups have become more influential.[29] In 2017, although 100 largest MNEs worldwide accounted for about 0,1 per cent of the total number of MNEs, their sales made up around 10 per cent of the world’s gross domestic product.[30]

The rise of corporate groups started to challenge the existing separate entity doctrine and a new policy was created:

[22] See Christians (2018, pp. 2–4).
[23] See Petrie (1952, pp. 7–9).
[24] See Ault, Arnold, and Cooper (2019), chapter “Canada".
[25] See Bunn and Asen (2019, p. 6).
[26] See Internal Revenue Service (2019b).
[27] See Canada Revenue Agency (2005).
[28] See OECD (2020).
[29] See Ting (2013, pp. 3–4).
[30] See United Nations Conference on Trade and Development (UNCTAD) (2018, p. 26).

the enterprise doctrine. This new provision implies that separate entities under one corporate group are treated as a single taxable unit. Following this change, countries worldwide reacted by adjusting their tax law. Currently, the responses vary in a large spectrum, whereby some jurisdictions apply relatively restrictive group taxation regimes and others employ more comprehensive consolidation schemes.

Group taxation regimes have two key functions: Intra-group loss offset and tax-free intra-group asset transfer. If desired, lawmakers can replace these two functions with full consolidation.[31] Since group taxation regime allows to reduce the tax burden by filing one single consolidated tax return, the existence and scope of group taxation regimes are attractive features of a country's tax environment.[32]

The group taxation regime of the U.S. was first introduced in 1917 as the primary form of group tax relief. The regime has experienced many changes over its long history and the policy has targeted companies that sought to decrease their taxes on their profits.[33] Section 1501 of IRC gives a group of corporations the right to opt in to file a single consolidated tax return to offset intra-group losses. An election to file a consolidated return needs pre-approval of the Internal Revenue Service commissioner.[34] In order to be qualified for the consolidation, the parent company must possess ownership of at least 80 per cent of both value and voting power of the subsidiary.[35] The primary advantage of this provision is to combine the income of the subsidiaries into a single taxable income figure. In addition, intra-group dividends can be ignored, and the recognition of the gains or losses of intra-group transactions can be deferred. Nevertheless, this provision only applies to domestic entities, whereby foreign incorporated subsidiaries must file a separate return. Exceptions exist for certain Mexican and Canadian subsidiaries. Following, problems arise for incorporated groups when faced with states' diverse and uncoordinated filing requirements. Since, while some states require separate filing for each corporation, others allow consolidation. In states where separate reporting is enforced, a corporation is obligated to determine separate income even though the entity is part of a consolidated group for federal tax purposes.[36] In addition, since partnerships are not corporations their return cannot be consolidated.[37]

On the other hand, there is no existing group taxation regime in Canada. Thus, differing from the U.S. taxation provisions, under Canadian ITA, each corporation is required to file a separate return.[38] The absence of group taxation provisions under Canadian law is a feature that reduces the attractiveness of the country.

[31] See Ting (2013, pp. 13-18, 32).
[32] See Schanz et al. (2017, p. 6).
[33] See Ting (2013, p. 70).
[34] See Brauner (2011, p. 70); Ting (2013, p. 301).
[35] See Trump (2020, p. 4).
[36] See Brauner (2011, p. 305).
[37] See PricewaterhouseCoopers International Ltd. (PwC) (2020b).
[38] See Dumalski and DeGrace (2019, p. 4).

3.2. Cost Recovery

3.2.1. Capital Cost Recovery

Investments promote growth in an economy; thus, tax systems have to be designed in a way that they are able to attract capital. Depreciation allowances play a major role when MNEs decide whether to invest in a jurisdiction.[39] Since in tax systems based on income an immediate write-off of the costs of capital investments is not allowed, businesses are required to spread the write-off of these costs over several years. The idea behind is that the legally allowed deductible amount reflects the depletion in value of an asset.[40]

However, write-offs in the future years prove as less valuable compared to earlier write-offs as a consequence of the decreasing time value of money. Consequently, firms cannot deduct fully the present value of the costs of their capital investments. Due to this decrease in the present value of depreciation allowances, true business costs are understated whereas the taxable income is being overstated. Overall, countries that offer faster write-off provisions become more attractive for MNEs, since their depreciation schedules serve to lower down the tax base easily.[41]

Capital cost recovery system of the U.S. is based on an asset classification system and it spreads out the depreciation costs over the estimated useful life of an asset.[42] Depreciation deductions may be taken for tangible properties that are used for business purposes or properties that produce income, such as rental property. The modified accelerated cost recovery system is available for properties placed after 1986, under which each class has a prescribed recovery period and depreciation method. General cost recovery periods depend on the type of tangible property and are three, five, seven, ten, 15, 20, 27,5 and 39 years (31,5 years for properties before 13 May 1993). Allowed depreciation methods are declining-balance method and straight-line method, switching from the point when the latter maximises the depreciation deduction. Taxpayers may switch to an alternative depreciation system and use only the straight-line over prescribed useful lives. Prior to the enactment of TCJA, a special 50 per cent first-year depreciation allowance applied, which has been replaced by the new regulation with a 100 per cent first-year bonus depreciation deduction for certain qualified properties that are acquired and placed in service after 27 September 2017 and before 1 January 2023. Tangible properties with a recovery period of 20 years or less are defined as qualified property under the modified accelerated cost recovery system, and special depreciation allowance reduces the basis before regular depreciation deductions.[43]

In Canada, similar to the U.S., capital cost allowances are calculated on the basis of pooled asset classes, whereby annual depreciation allowances are given as a prescribed rate

[39] See Nam and Radulescu (2005, p. 5).
[40] See Brazell and Mackie (2000, pp. 532–533).
[41] See Bunn and Asen (2019, pp. 8–9).
[42] See Brazell and Mackie (2000, p. 531).
[43] See Ernst & Young LLP (2019b, p. 1818); PricewaterhouseCoopers International Ltd. (PwC) (2020a).

and applied on a declined-balance basis. Businesses cannot deduct the full amount in the year of acquisition of depreciable property and have to disperse the capital costs over several years. In response to TCJA, in the Fall Economic Update of 2018, the federal government of Canada introduced three important changes with regard to capital cost allowances to "enhance business confidence".[44] According to the new provisions, the cost of machinery and equipment used for manufacturing or processing of goods may be immediately written-off. Since the manufacturing industry is capital intensive and highly mobile, this new incentive will help the industry to cope with the changes brought by TCJA. Secondly, businesses are given the chance to immediately write-off the full costs of clean energy equipment. Furthermore, the federal government introduced the Accelerated Investment Incentive, under which businesses can deduct their capital investment costs at an accelerated rate. This accelerated rate is approximately equal to three times the usual first-year deduction and it will gradually phase out by 2028.[45] The temporary accelerated deduction provisions of both countries increased the allowed amount of the first-year deductions. In a way, the Fall Economic Update of Canada has replicated capital cost recovery provisions of TCJA to establish a similar attractiveness for capital investments as the U.S. in its jurisdiction.

3.2.2. Loss Carryforward

Companies occasionally have unused tax credits in a given tax year which is mostly the case in times of economic downturn. Since tax carryforwards imply future tax savings, they are valuable assets to the firm. In most countries, companies are allowed to offset their current losses against future profits. In other words, if the company has a positive tax base in the future, it can deduct its tax loss carryforwards to reduce its tax burden.[46] Hence, the amount of a tax loss carryforward is a considerable contribution to the company's value[47] and it helps to "shelter a part of the firm's future income from taxation".[48] Loss carryforward provisions of a country dictate the number of years a company is permitted to carry forward its losses. Countries with more generous provisions are more attractive to MNEs.[49]

According to the U.S. IRC, net operating losses (NOL) arise when deductions exceed gross income in a given tax year. NOLs generated before 1 January 2018 may be carried forward for up to 20 years. TCJA updated this provision, and currently, NOLs generated after that date can be carried forward indefinitely. Nevertheless, for NOLs without carryforward limitations, the deduction is capped to 80 per cent of the taxable income which has been calculated without regard to the deduction.[50] Other limitations for the carryforward of NOLs are defined in Section 382, whereas special rules apply for deductions subsequent to a reorganisation and after changes in corporate ownership. On 27 March 2020, the Coronavirus Aid, Relief, and Economic Security (CARES) Act was enacted which provided relief from relatively strict provisions of TCJA. One of those provisions was the elimination of the income limitation of NOLs for carryforwards for post-TCJA years through 2020.[51]

Different from the U.S. provisions, ITA of Canada differentiates among capital losses, allowable business investment losses, non-capital losses and farm losses. Non-capital losses are realised from regular business activities and they can be deducted against any source of income. The unused portion of the non-capital losses can be carried forward for up to 20 years.[52] Similar to the U.S., after the prescribed time period the rights to offset losses against income becomes permanently unavailable. Much like the provisions of IRC, Canadian lawmakers have limited the carryforward option in case of an ownership change. After a majority acquisition loss carryforward can only be deducted if the acquired business will continue to operate with reasonable profit expectations. Also, deduction is allowed in cases where the losses are used to offset the same or similar business income. The idea behind this loss streaming rule is to avoid acquisitions that solely aim to utilise loss carryforwards of acquired companies.[53]

3.2.3. Loss Carryback

Loss carryback allows companies to deduct current year losses against past year profits. Analogously to loss carryforwards, loss carrybacks reduce a company's overall tax burden. Thus, countries that allow companies to carry back their losses far into the past are more attractive from the tax perspective.

In the U.S., prior to TCJA, NOLs were permitted to be carried back for two years to obtain a refund. TCJA changed this rule and eliminated the loss carryback provision completely.[54] Nevertheless, this provision was recently modified with the CARES Act, whereby corporate taxpayers are now permitted to carryback their NOLs for the years 2018, 2019 and 2020 for up to five years. Taxpayers have the option to choose one particular tax year for carryback. However, in general, once a carryback is claimed, then it must be executed for all five years. As a result, it is expected that loss carrybacks introduced with the CARES Act will provide economic relief for companies that are affected by the Coronavirus pandemic.[55]

On the other hand, differing from the U.S. provisions, non-capital losses may be carried back for three years in Canada to offset gross income arising in those tax years. In the event that the carryback losses have not been utilised

[44] Department of Finance Canada (2018, p. 55).
[45] See Ernst & Young LLP (2019a, p. 266); PricewaterhouseCoopers International Ltd. (PwC) (2019).
[46] See Streitferdt (2013, pp. 163–164).
[47] See Betker (1995, pp. 3–18) as cited in Streitferdt (2013, p. 12).
[48] Sarkar (2014, pp. 803–804).
[49] See Bunn and Asen (2019, p. 8).
[50] See PricewaterhouseCoopers International Ltd. (PwC) (2020a).
[51] See Singer, Bray, Kershaw, Shulman, and Crouch (n.d.).
[52] See Canada Revenue Agency (2017).
[53] See Thomson Reuters Canada Ltd. (n.d.).
[54] See Internal Revenue Service (2019a).
[55] See Luscombe (2020, p. 3).

in the prescribed time frame, they become permanently unavailable to the taxpayer.[56] Compared to the U.S., Canada has not yet introduced any relief provisions for business taxes with regard to losses encountered due to the pandemic. Despite the fact that the CARES Act introduces provisions limited to a short time period, it is still important when considering the tax attractiveness since it shows the U.S. government's determination on supporting its businesses.

3.3. Tax Incentives

3.3.1. Research and Development

Tax incentives used to promote research and development (R&D) business activities are adapted widely among many countries and they are a tool for national growth and job creation.[57] It has the characteristic of reducing the costs of carrying out R&D activities by alleviating MNEs' tax liability. Since R&D investments are large expenses, tax incentives that reduce the after-tax R&D costs are important when determining the tax attractiveness of a jurisdiction.[58] The economic rationale is derived from the statement that social returns of inventions are higher than the private returns. In other words, the end-product of an R&D activity can be appropriated in a private enterprise to a limited extent compared to public purposes. Thus, companies would be reluctant to invest in R&D activities without any incentives.[59]

The policy design of tax incentives of R&D activities can be volume-based or incremental, whereby the simpler design is, the higher the welfare is going to be due to the alleviated administrative burden. The U.S. and Canada both employ a volume-based policy, whereby the total amount of qualified R&D activities is taken into account. Furthermore, tax incentives covered by R&D are tax credit, enhanced deductions and depreciation allowances for R&D activities. The tax credit allows firms to directly deduct the amount determined by their tax rate from their tax liability born by their R&D expenditures. The U.S. and Canada both implement the method of granting tax credit as a way to incentivise R&D activities.[60]

In the U.S., in addition to the federal tax credit, many states offer separate R&D tax credits that are generally at a lower level than the federal credit rate. The traditional research tax credit is 20 per cent of the qualified research expenses that exceed the pre-defined base amount whose calculation relies on data between 1984–1988. Due to the complexity of defining the base amount, only a few companies choose to report the traditional research credit. Alternative simplified credit is another option that is more widely used, under which the credit rate is 14 per cent of the excess expenses, whereby the threshold is 50 per cent of the averaged past three consecutive years' qualified research expenses. Lastly, the U.S. government offers targeted research credits that, as the name suggests, focus on specific types of R&D activities. MNEs that qualify for targeted research credit cannot use the regular research credits for the same qualified expenses. The R&D tax credit system of the U.S. does not exclude any industries, thus all companies that undertake qualified research expenses are eligible for a tax credit.[61]

Canada has a relatively lucrative tax credit system compared to other countries around the world and it is organised under the Scientific Research and Experimental Development (SR&ED) Tax Incentive Program. SR&ED federal investment tax credit is 15 per cent for all qualifying R&D expenditures carried out in Canada, and it is used to reduce federal tax payable. On the other hand, different from the U.S., Canada does not offer many diverse credit options besides an additional enhanced refundable tax credit of 35 per cent for Canadian controlled private corporations on eligible expenditures of three million Canadian dollars. For R&D expenditures that exceed this amount, there is also the option to earn a non-refundable 15 per cent tax credit.[62]

Qualified expenses under the U.S. R&D tax credit system include wages and salaries of employees and all types of supplies other than depreciable assets. On the other hand, Canada only takes into account materials that are used for the purposes of the research. As a result, the U.S. IRC qualifies more items as eligible for R&D tax credit purposes since "supplies" cover a wider range of items than "materials", therefore it gives companies more discretion in determining the activities that constitute as R&D expenditure.[63]

Additionally, similar to the U.S. state tax credit, Canada has separate provincial/territorial credits that range from 3,5 per cent to 20 per cent. MNEs both in the U.S. and Canada are allowed to carry forward their tax credits for 20 years and claim them back in future periods. SR&ED of Canada also offers an immediate deduction for all qualified expenditures, under which there is the option to carry forward and deduct in future years.[64]

3.3.2. Patent Box Regime

Historically, R&D tax incentives have been seen as the only effective way of promoting innovative business activities. However, in the early 2000s, the current form of the patent box regime has emerged, and it has been found that incentives subsidising the income stream from innovation activities result with more favourable outcomes in the long run than incentives that subsidise investments in innovation, i.e. R&D tax incentives. Patent box regime has the objectives of stimulating domestic innovation and combatting erosion of the domestic tax base. It reaches these objectives by featuring a reduced tax rate on income derived from the commercialisation of royalties or giving a tax exemption on a certain percentage of the latter.[65] The existence and scope of a patent

[56] See Canada Revenue Agency (2017).
[57] See Goulding and Goulding (2013, p. 18).
[58] See Schanz et al. (2017, p. 8).
[59] See Arrow (1962, pp. 616–619).
[60] See Correa and Guceri (2013, pp. 5–6, 10–11).
[61] See Aronshtam and Kane (2018, pp. 271–273).
[62] See Canada Revenue Agency (2020).
[63] See Laplante, Skaife, Swenson, and Wangerin (2019, pp. 89–91); SRED (2017).
[64] See Aronshtam and Kane (2018, pp. 35–37); OECD (2019, p. 1).
[65] See Bradley, Dauchy, and Robinson (2015, pp. 1047–1048, 1069).

box regime increase the tax attractiveness of a jurisdiction substantially.

Prior to the introduction of BEPS project Action 5[66], the policymakers in the U.S. have not been showing serious interest in the development of patent/innovation box regimes. However, with the introduction of the modified nexus standard, the U.S. government felt obliged to respond to hinder redirection of R&D activities by U.S. MNEs. In July 2015, the Innovation Promotion Act[67], publicly known as the Boustany-Neal discussion draft was released. This Act amended IRC by adding new provisions related to the innovation box and it was stating that the enactment of U.S. innovation box would allow to increase the competitiveness of U.S. companies, but this Act has never passed the Senate.[68]

Until the enactment of TCJA, no further discussions related to the discussion draft took place. Although TCJA does not include any direct patent/innovation box regimes, it introduced two special provisions: FDII and GILTI. According to FDII provision, 37,5 per cent of the income derived from the commercialisation of the licences of intellectual property is subject to deduction. Nevertheless, on the opposite side, GILTI provision imposes a minimum tax on the global income, including income derived from licencing of intellectual property. Thus, some commentators argue that FDII and GILTI act as “carrot and stick", and only create an indirect patent/innovation box regime.[69]

As was the case with the U.S., Canadian policymakers have been discussing a possible patent box regime for years, but it has not yet been implemented on the federal level. However, regimes on a provincial level have been adopted. British Columbia was the first province in Canada who granted a reduced tax rate of 2,75 per cent for international patent income. The province does not enforce the product to be developed locally, whereby companies with a permanent establishment in the province are entitled of the tax rate reduction. The second provincial patent box regime was then introduced in Quebec in 2016 to bridge the gap between innovation and its commercialisation, where the tax on qualifying patent income was reduced to a rate of 4 per cent. However, Quebec’s patent box regime only applied to the manufacturing industry and it reduced the tax rate for qualified innovative manufacturing corporations, whereby intellectual property has to be at least partly developed in Quebec.[70]

3.4. International Tax Regulations

3.4.1. General Anti-Avoidance Rules

Taxpayers may engage in tax planning strategies to reduce their tax payable. Notwithstanding, in cases, where tax avoidance strategies have reached a drastic level, governmental interference becomes necessary. Therefore, GAAR are implemented by many jurisdictions to restrict impermissible tax avoidance.[71] The definition of the latter can become problematic because of its ever-changing and unpredictable nature; however, it is broadly characterised as being inconsistent with the spirit of tax law since these transactions are often abnormal, artificial and they lack a commercial substance. Furthermore, there is an additional complexity born with GAAR, since there is a clash between taxpayers’ entitlement to lawful tax planning, and the governments’ need to protect its tax revenue from impermissible tax avoidance.[72] On the other hand, for MNEs the existence of GAAR provisions are a drawback since they prefer having freedom in their tax planning strategies. A country that imposes clearly defined, strict GAAR rules is not considered as attractive compared to other jurisdictions without any anti-avoidance provisions.

In the U.S., in accordance with the common law tradition at the federal level, the judiciary decides ultimately, whether there is a tax avoidance transaction and whether resulting tax benefit from this transaction should be denied. The anti-avoidance doctrine was first developed with the case Gregory v. Helvering[73] and stated that economic substance of a transaction is the deciding factor in its classification. Lack of economic substance signals that a business has been solely transacted with the purpose of tax minimisation.

In 2010, the economic substance doctrine was codified with the enactment of Healthcare and Reconciliation Act in IRC Section 7701(o).[74] Nevertheless, the codified version has drawn criticism with the argument that it “provided a roadmap to successful avoidance"[75] compared to the prior judicial version. Thus, despite the codification of the doctrine, federal courts still hold the power to determine whether a transaction is part of a tax avoidance strategy or has an economic substance within it.[76] In addition to the economic substance doctrine, U.S. courts rely on other doctrines such as substance over form, step transaction and sham transaction doctrines.[77]

In Canada, taxpayers’ general right to arrange its affairs in a way that its tax burden is reduced, originated from the “Duke of Westminster principle"[78], under which legitimate right to engage in tax planning has been recognised by the Canadian courts. However, the Stubart[79] case has shown the need for a general provision aimed to limit abusive tax avoidance and following, a robust GAAR provision has been enacted in 1988. The Canadian anti-avoidance provision has been defined in Subsection 245 of ITA.

Different from the U.S. GAAR, Canadian GAAR works un-

[66] OECD (2015).
[67] Goodlatte (2015).
[68] See Atkinson (2015, pp. 5–8); Traub and Kummer (2015, pp. 3–4).
[69] See Shaviro (2018).
[70] See De Luca et al. (2016, p. 3).
[71] See Waerzeggers and Hillier (2016, p. 1).
[72] See Kujinga (2012, pp. 42–43).
[73] Gregory v. Helvering (1935, para. 2).
[74] See Prebble and Prebble (2010, p. 27).
[75] See Avi-Yonah and Pichhadze (2017, p. 10).
[76] See Brown (2012, pp. 609–610).
[77] See Larsen, Cordero, Bowen, Denault, and D’Alessandro Jr. (2018, p. 21).
[78] Duke of Westminster v. Commissioners of Inland Revenue (1935).
[79] Stubart Investments Ltd. v. The Queen (1984).

der a cascade mechanism which gives little leeway to taxpayers. Firstly, Subsection 245(2) determines whether there is a tax benefit resulting from the transaction. If there is one, it is then to decide whether the transaction is motivated for tax purposes or *bona fide* non-tax reasons according to Subsection 245(3). Lastly, Subsection 245(4) examines to see if tax benefit resulted from a misuse of specific taxing provisions of ITA. As is the case with the U.S., subjective nature of GAAR is reflected in the Canadian provisions, thus statutory interpretation by the courts is an essential element.[80] In addition to GAAR provisions, the U.S. and Canada have also specific anti-avoidance rules that go beyond the general provisions and apply to narrowly prescribed set of transactions[81], whereby these rules will be examined separately in the coming parts.

3.4.2. Controlled Foreign Corporation Rules

Global trade landscape is dominated by a high level of global integration which complicates the taxation of corporate income and makes it a border-crossing issue. In cases where profits are generated both in the home country of the parent company and in the host country of a subsidiary, it becomes problematic in which jurisdiction this income should be taxed. Each entity of the corporate structure is regarded as a separate taxpayer; thus, each subsidiary becomes subject to income tax in the host country.

Following, most countries impose domestic income tax when the income has been repatriated and allow methods to eliminate double taxation that arises from taxing the same base multiple times. Nevertheless, the parent company owns essentially the profits and hence has the right to determine when, if at all, the income should be repatriated to the home country. Considering that base protection and enforceability of tax system are integral parts of a country's optimal policy design, when profits escape the reach of the home country, it possesses a threat to the enforceability of tax collection.[82]

In order to control aggressive tax planning and constant migration of profits, many high-tax countries introduced CFC rules. These rules impose a direct tax on foreign subsidiaries by including non-repatriated income into the domestic tax base. Whether to enforce CFC rules or not depends generally on three criteria: Control criterion, tax level criterion and source of income. Control criterion requires the shareholder to be in control of financial decisions of the entity in a foreign country and the ability to make decisions related to tax avoidance. On the other hand, the tax level criterion obliges a fixed minimum tax level, below which incentives for profit shifting emerge. Lastly, the source of income should be passive income, whereby any highly mobile income that makes income-shifting possible, qualifies, e.g. foreign personal holding company income. Since the CFC rules greatly increase the total tax burden of MNEs, countries that impose these rules are seen as less attractive for investment purposes.[83]

The U.S. was the first country worldwide who introduced CFC rules in 1962. It took over a decade for other capital-exporting OECD countries, including Canada, to adopt similar rules.[84] Under U.S. law, CFC is a foreign corporation that is owned by over 50 per cent by U.S. shareholders, whereby certain types of income are included in the taxable income of the U.S. shareholders. Types of income that must be included in the tax base are defined in subpart F of IRC which includes foreign-based company sales income and foreign-based company service income, among others. U.S. shareholders are by definition, U.S. persons who own more than 10 per cent of a foreign corporation's voting stock or own more than 10 per cent of the total value of all classes of a foreign corporation's stock.[85] Thus, different from Canada, the U.S. implements a two-step ownership test, whereas first, the ownership of the corporation is examined, and afterwards the shareholders' ownership is assessed. Following, the dividends of the CFC received by U.S. shareholders are eligible for a full "dividend-received deduction", which means that dividends obtained are tax-free since CFC did not make any tax payments to the government.[86]

Nevertheless, after the enactment of TCJA, IRC Section 951 Subtitle A is introduced, and it required U.S. shareholders to include the amount of GILTI, independent of whether the amount is distributed or not. As a result, TCJA's GILTI rules expanded the scope of CFC rules beyond passive income, thus it is now applicable to all income generated by a CFC if the income is considered to be low-taxed. In other words, TCJA does not define any tax havens, it rather taxes all income that is subject to a foreign tax rate lower than 13,125 per cent. GILTI is despite its name, is not limited to low-taxed income from intangible assets, and encompasses U.S. shareholders' pro-rata share of CFC's total net income less 10 per cent return on depreciable tangible property.[87] GILTI is included in the U.S. shareholder's income and a deduction equal to 50 per cent of the full amount is allowed. In addition, the shareholder has the right to claim a foreign tax credit for 80 per cent of the foreign taxes paid or accrued by a CFC.[88]

Another provision that was enacted with TCJA was the Base Erosion and Anti-Abuse Tax which imposes a minimum tax to the tax base of large corporations with gross receipts of 500 million dollars or more, and in addition to corporations that have deductions that exceed 3 per cent of their total deductions paid to foreign related companies. Base Erosion and Anti-Abuse Tax, as the name suggests, levy a direct tax on base-eroding payments and acts as a "backup to GILTI".[89]

[80] See Hashimzade and Epifantseva (2017, p. 108).
[81] See Marley and Tremblay (2005, p. 105).
[82] See Clifford (2019, pp. 44–46).
[83] See Schanz et al. (2017, p. 5).
[84] See Arnold (2012, p. 478).
[85] See Traub and Kummer (2015, p. 8).
[86] See Yang, Poon, and Chang (2018, p. 20).
[87] See McGill, Karch, Feeley, O'Banion E., and Crouse (2018, pp. 11–13).
[88] See PricewaterhouseCoopers International Ltd. (PwC) (2020b).
[89] See Dueñas (2019, p. 18).

Canadian CFC rules are organised extensively under the Subsection 91(1) of ITA and named as foreign accrual property income rules.[90] Similar to the U.S. provisions, foreign accrual property income regime obliges a Canadian resident shareholder to include its share of passive income of its controlled foreign affiliate, whereby it is irrelevant whether or not this income has been distributed. In order to be qualified as a controlled foreign affiliate, the majority of the voting shares must be owned directly or indirectly by a Canadian corporation. Foreign accrual property income covers income from investment business and income that is derived from the trading or dealing in indebtedness. In case of income is derived from specific activities, it encompasses the income if it accrues to the benefit of the Canadian taxpayer under a tracking arrangement.[91]

3.4.3. Thin Capitalisation Rules

The trade-off theory of corporate finance states that firms pick their capital structure by trading off the benefits of tax shield from debt against its costs, i.e. financial distress and agency costs. A tax shield is a reduction in tax expenditures due to tax-deductibility of interest expenses.[92] In addition to the advantages raised by this asymmetry of tax treatment, differences in tax rates across countries result with profit shifting and reduction of the overall tax burden.

MNEs allocate their debt to high-tax countries or provide loans internally to related entities in those jurisdictions in order to deduct higher interest expenses from their tax base.[93] Consequently, entities in low-tax countries are equipped with equity, whereas entities in high-tax countries have large amounts of debt on their balance sheets. In order to limit excessive debt financing and intra-group financial transactions, lawmakers in high-tax countries have adopted thin capitalisation rules. These rules aim to prevent profit shifting to a large extent by limiting the deductibility of interest payments from taxable income. Thin capitalisation rules differ greatly across countries and usually have a high degree of complexity. They can also be arranged in a way that they are only binding for certain selected industries.[94] Since interest tax shield is an important financial instrument to reduce tax payable of corporations, provisions that limit interest deductions are not favoured by MNEs. Hence, they reduce the tax attractiveness of a country.

The U.S. government imposes clearly defined thin capitalisation rules which have been tightened with the enactment of TCJA. The current allowed debt-to-equity ratio is 1,5:1. Section 385 of IRC contains provisions that evaluate whether an instrument claimed to be debt is considered as debt or equity for the U.S. tax purposes. If the transaction is legally prohibited, capital loaned to an entity by a related party may be characterised as equity by Internal Revenue Service. Correspondingly, its interest expense deductions may be disallowed, and its principal repayments and its interest expenditures are considered as distributions and they are subject to withholding tax.

If the instrument is considered as debt, the business interest expense deductions apply.[95] These deductions have been modified with the enactment of TCJA and they have been limited to 30 per cent of the adjusted taxable income. Deductions apply broadly to the business interest of any taxpayer; hence it is irrelevant whether they are foreign or U.S. persons or whether they are related or unrelated entities. For tax years before 2022, the adjusted taxable income equals approximately to earnings before interest, taxes, depreciation, and amortisation, whereby after that date it will be decreased by depreciation and amortisation. Carry forward of disallowed interest expenses into future years is allowed. In case of lending from a foreign lender, interest expenses must be paid prior to performing their deduction.[96]

As stated in Subsection 18(4) of ITA, Canada's thin capitalisation rules, rely on mandatory debt-to-equity ratios to prevent illegitimate diversion of profits to low-tax jurisdictions.[97] These provisions have been subject to amendments over the years. According to Section 385 provisions, Canadian lawmakers give a specified debt-to-equity ratio. Currently, Canadian thin capitalisation rules restrict the ability to deduct interest expenses paid by a resident corporation to a specified non-resident shareholder or related persons on debts exceeding the defined debt-to-equity ratio of 1,5:1. A specified shareholder is someone, who owns 25 per cent or more of a corporation resident in Canada, either by fair market value or by voting rights, with other related persons that do not deal at arm's length.[98] These rules apply to corporations resident in Canada, trusts, non-residents' Canadian branches and partnerships of which a corporation resident in Canada is a member. In addition, thin capitalisation rules apply to certain back-to-back loan arrangements.[99]

3.4.4. Transfer Pricing Rules

MNEs manipulate transfer prices when they conduct transactions with related companies in a way that their tax liabilities are minimised. They do so by setting high prices for the sale of goods or services to related entities in high-tax countries. As a result, they reduce the overall tax burden of the organisation and repatriate higher profits to low-tax countries. Hence, from the standpoint of MNEs non-existing or relatively vaguely organised transfer pricing rules increase the attractiveness of a jurisdiction.

Notwithstanding, for governments in cases when there are no existing restrictions with regard to the prices set for

[90] See Nosikova (2015, p. 4).
[91] See Dumalski and DeGrace (2019, p. 3); PricewaterhouseCoopers International Ltd. (PwC) (2020b).
[92] See Berk and DeMarzo (2017, pp. 552–556).
[93] See Merlo and Wamser (2014, pp. 27–28).
[94] See Schanz et al. (2017, p. 10).
[95] See Ernst & Young LLP (2019b, pp. 1821–1822).
[96] See PricewaterhouseCoopers International Ltd. (PwC) (2020a).
[97] See Nitikman (2000, pp. 20–21).
[98] See Shafer (2013, p. 1).
[99] See Ernst & Young LLP (2019a, pp. 267-268).

intra-group transactions, transfer pricing becomes a sole device of tax evasion. Furthermore, tax authorities try to recover this revenue loss by restricting the setting of transfer prices by MNEs. Nevertheless, this intrusion of tax authorities may cause inefficiencies that damage consumer prices and MNEs' organisational choices.[100] An intrusion as such also distorts the main point of MNEs' cross-national activity which is to cut costs by locating their organisation more efficiently around the globe.[101] In order to solve these conflicting problems, OECD has suggested a set of guidelines[102] to help safeguard tax authorities interests while alleviating market distortions raised by abusive transfer pricing strategies. Transfer pricing rules worldwide are organised based on the arm's length principle, under which the price that would prevail between two unrelated entities in a competitive market is predicated.[103]

The primary legislation of the U.S. transfer pricing rules is defined in Section 482 of IRC (Regulations 1.482-1 to. 1.482-9), also known as Treasury Regulations. It ensures that controlled transactions' profits are correctly reflected, and a controlled taxpayer is on par with an uncontrolled taxpayer. Along establishing a general standard applicable to transactions between related entities, Section 482 also introduces an additional standard for transfers of intangible properties. Although the arm's length principle has not been explicitly stated, it has been embodied by the statutory language by establishing a tax parity between controlled and uncontrolled transactions. According to Section 1.482-1 of Treasury Regulations the "best method rule" should be used to ensure the arm's length pricing, insomuch as the best transfer pricing method is the one with the most reliable arm's length result.[104] This method should be based on the criteria of comparability of uncontrolled transactions and quality of data and assumptions. Acceptable transfer pricing methods in transfers of tangible properties are comparable profits method, comparable uncontrolled price method, cost plus method, among others. Unspecified methods may also be used by taxpayers if they end in with an arm's length result. Furthermore, the U.S. rules for transfer pricing do not predicate on the guidelines issued by OECD and there are no references to the latter in Treasury Regulations. Advance pricing agreements can be reached between taxpayers and the Internal Revenue Service, under which an arm's length pricing method is agreed upon for the upcoming intercompany transaction.[105]

Canada's status as being a high-tax jurisdiction has become stronger after the corporate tax rate reduction in the U.S., which also caused transfer pricing rules gaining importance in the Canadian taxation system. Transfer pricing rules are contained in Section 247 of ITA, whereby, as is the case with the U.S., OECD's guidelines play a limited role for Canada's transfer pricing provisions. As stated in GlaxoSmithKline Inc. case[106], OECD's provisions may only be used as interpretative aid and only Section 247 of ITA is legally binding. Following, although, Section 247 does not prescribe any particular method, it applies the arm's length principle by comparing controlled transactions to comparable arm's length transactions. However, ITA does not specify any particular method, thus "the selection of the most appropriate pricing method depends [...] on the assessment of the comparability of transactions".[107] Acknowledged methods are similar to the ones recommended in the OECD's guidelines and include comparable uncontrolled price method, resale price, profit split method, among others. Although all methods are accepted, in the past courts have shown a preferential position to traditional transaction methods, specifically comparable uncontrolled price method. As in the U.S., advance pricing agreements with the Canada Revenue Agency are possible to provide more tax certainty for taxpayers.[108]

3.4.5. Treaty Network

Double taxation treaties play a crucial role as a policy tool to promote cross-national investment activity.[109] These treaties have two primary objectives: avoiding double taxation in international transactions and preventing tax evasion by MNEs. There is a consensus that double taxation treaties are mean to reduce withholding taxes on passive income. Since they eliminate to a large extent the investment uncertainty that a company may encounter when investing overseas, they also help to promote foreign direct investments.[110] As a result, an increase in foreign direct investments spurs economic growth and hence raises the tax revenues of governments. Thus, the number of effective tax treaties that are in force boosts the attractiveness of a country for MNEs. Double taxation treaties are largely influenced from the standards defined by the OECD Model Tax Convention on Income and on Capital[111] and United Nations Model Double Taxation Convention[112], and they shift the taxing rights from the source state to the resident state.[113]

The U.S. receives the biggest share of worldwide foreign direct investments due to the relative openness of its economy. A vast majority of the U.S. foreign direct investments comes from the countries with which the U.S. has signed a tax treaty in the past. Currently, the U.S. has 64 bilateral tax treaties in force.[114] In the past, the U.S. had viewed its treaties as negotiated agreements with strong bilateralism. Nevertheless, the U.S. government has decided in 1976 to

[100] See Behrens, Peralt, and Picard (2014, pp. 651–652).
[101] See The Economist (1993, para. 8).
[102] OECD (2017b).
[103] See Taklalsingh (2019, pp. 3–4).
[104] See Ernst & Young LLP "Worldwide Corporate Tax Guide" (2019), p. 1822.
[105] See Larsen et al. (2018, p. 11).
[106] Canada v. GlaxoSmithKline Inc. (2012, para. 20).
[107] Alberta Printed Circuits Ltd. v. The Queen (2011, para. 167).
[108] See Suarez (2019, pp. 797–799).
[109] See Kunka, Stasio, and Zagler (2020, pp. 575–576).
[110] See Daniels, O'Brien, and von der Ruhr (2015, p. 1000).
[111] OECD (2017a).
[112] United Nations Model Double Taxation Convention between Developed and Developing Countries (2017).
[113] See Braun and Fuentes (2016, pp. 384–385).
[114] Deloitte (2020).

Criteria	Differences
Corporate Income Tax Rate	Federal tax rate in the U.S.: 21 per cent Combined average tax rate in the U.S.: 25,77 per cent Federal tax rate in Canada: 15 per cent Combined average tax rate in Canada: 26,47 per cent
Group Taxation Regime	There is no group taxation regime in Canada.
Capital Cost Recovery	Different rates for temporary accelerated deduction provisions.
Loss Carryforward	Carryforward in Canada is allowed for 20 years. In the U.S carryforward allowance is indefinite but is capped to 80 per cent of the taxable income.
Loss Carryback	Canada allows carryback for three years. No carryback allowance in the U.S.
Research and Development	Eligible R&D expenditures are more broadly defined in the U.S. than Canada.
Patent Box Regime	Two provinces in Canada have employed patent box regimes. In the U.S. there is an indirect patent box regime with FDII and GILTI.
General Anti-Avoidance Rules	GAAR provisions of Canada work under a cascade mechanism. The U.S. GAAR relies on many doctrines.
Controlled-Foreign Corporation Rules	GILTI provision expands the scope of eligible income beyond passive income. Base Erosion and Anti-Abuse Tax of the U.S. acts as back-up to the CFC rules.
Thin Capitalisation Rules	The U.S. limits interest deductions to 30 per cent of adjusted taxable income, whereby there are no limitations in Canada. Tighter thin capitalisation rules in the U.S.
Transfer Pricing Rules	N/A
Treaty Network	Canada has a larger number of bilateral tax treaties than the U.S.

Figure 1: Brief overview of the differences

publish a draft “U.S. Model" which is essentially based on the OECD's Model, and hence it is very similar in nature. However, there are few substantial deviations such as corporate residence. In the U.S. tax treaties, the latter is determined solely based on its place of organisation, whereas this is not the case in OECD's Model and in the United Nations' Model. These deviations are a reflection of the uniqueness of the U.S. tax treaty provisions.[115]

The treaty network of Canada has grown tremendously since 1976 when the link between treaties and Canada's exemption system for the taxation of active foreign-sourced income has been established. According to this provision, in case of profits are generated in a designated treaty nation, Canadian MNEs may repatriate their foreign-sourced active income to Canada tax-free. Different from the U.S. treaty network, all the Canadian tax treaties are based on the OECD's Model and on the United Nations' Model. However, specific economic relations with a particular country are reflected individually in each treaty.

Compared to the U.S., Canada has a larger treaty network. Currently, the country has 93 bilateral treaties in force and four additional treaties that are signed but not yet in force.[116] Among existing tax treaties of Canada, the most important one is the treaty with the U.S.[117] The rationale behind this statement derives from the strong economic links between two countries; whereas the U.S. receives the biggest share of foreign direct investments of Canada.[118]

4. Lessons for Canada from the U.S.

Historically, Canada provided a more attractive tax environment for MNEs compared to the U.S., thus the country enjoyed an inward shift of foreign corporations' profits. Nevertheless, this tax advantage faded with the enactment of TCJA. While it is not yet certain how large the consequences of the reform are going to be, it is a certain fact that Canada's tax attractiveness started diminishing. The measures that were introduced with the Fall Economic Update in response to TCJA were temporary in nature. Hence, a robust and internationally competitive reform is necessary for Canada to regain its position in global trade scenery.[119]

When preparing a reform as such, Canadian lawmakers could look at the elements of the U.S. tax system. This part of the thesis derives lessons for Canadian government from the results of the comparison in the previous section. The results of the comparison indicated that among these 12 criteria, all of them except “transfer pricing rules" can be identified as different, nonetheless, the degree of divergences varies. Two criteria, group taxation and loss carryback, differ the greatest since while one country has the respective provision whereas

[115] See Brauner (2017, pp. 21.23).
[116] See Department of Finance Canada (2019).
[117] United States - Canada Income Tax Convention (1985).
[118] See Cockfield (2017, pp. 118–119).
[119] See McKenzie and Smart (2019a, pp. 2–4, 22).

the other has not. Broadly speaking for five[120] out of 12 criteria the differences are of disadvantage for Canada. In this part, in addition to the results of the comparison, the views of commentators from governmental agencies and academia are included. Since the commentators also formulated their suggestions while taking the U.S. system into account, the lessons that were derived from the previous part of this thesis and their views are harmonised to a large extent.

As indicated earlier, lower corporate income tax rate makes a country more attractive and draws capital investment. In this regard, Canada has lost its competitive edge over the U.S. with the reduction of the U.S. corporate income tax rate. Moreover, it has been found that companies in non-tax haven jurisdictions are more sensitive to movements in corporate tax rate which makes a Canadian response even more necessary.[121] However, critics' views vary with regard to the amount of reduction in corporate income tax rate whereas some do not recommend turning this to "a race to the bottom"[122], and others suggest a reduction of about 10 per cent.[123] Canadian policymakers should weigh the pros and cons of both approaches, nevertheless, the need to respond to the drastic reduction of the U.S. tax reform stays further on necessary.

Canada currently has no group taxation regime. Implementing a group taxation regime would move some MNE activity to Canada since it allows corporations to reduce their overall tax burden by transferring losses and filing a single consolidated tax return.[124] However, the case of the U.S. shows that when adding group taxation provisions into its law, Canadian lawmakers should also consider their negative aspects: The possibility of consolidation cleared the way for aggressive tax planning strategies in the U.S. which gave rise to the need of specific anti-abuse provisions.[125] Consequently, group taxation regime is a double-edged sword, hence when implementing similar provisions to the ones in the U.S., policymakers must work thoroughly to balance its positive and negative aspects.

As stated above in section 3.2.1, the Canadian government announced temporary accelerated capital cost recovery provisions to raise capital investments as a response to the measures implemented with TCJA. It has been argued however that these temporary measures are inadequate, and they make already bafflingly organised ITA of Canada even more complicated.[126] Thus, when moving forward, instead of implementing further interim measures, the Canadian government should prefer more grounded changes. A proposal is to implement a "cash-flow tax" that taxes the economic rents of an entity. This model would allow immediate deduction of all capital expenses and eliminate the interest deductions. As a result, all sectors will be influenced equally and distortions between different asset classes will be eliminated.[127]

In case of loss offset rules, the time value of money enhances the importance of the number of time periods for which the carryover is allowed. Since loss carrybacks are retrospective, the losses can be fully utilised. Notwithstanding, loss carryforwards can only be partially utilised because of the declining time value of money. As a result, the longer a company carries forward its losses, the more the value of the latter decreases.[128] Thus, an increase in Canada's carryforward allowance in years similar to the increase in the U.S. is not expected to have a major impact. Canada already has a relatively long carryforward time period which can be considered as attractive.[129]

As discussed above in section 3.3, tax incentives are important means to promote innovation in a country. SR&ED program of the Canadian government provides good incentives for R&D activities; however, innovation activities of Canadian corporations are still considered as inadequate. Research suggests that many corporations leave the country before they grow and thus do not commercialise their intellectual properties in Canada.[130] In order to change this phenomenon, federal measures are necessary. For instance, a patent box regime on the federal level would encourage companies to engage in R&D activities and also increase the attractiveness of Canada.[131] Another direct solution would be to increase the funding of SR&ED program. Furthermore, the Canadian government could ensure corporations that are looking to expand to new jurisdictions that their intellectual property rights are going to be protected in these countries. This can be provided by assuring increased protections in trade agreements.[132]

With regard to GAAR provisions, there are not many significant differences between the U.S. and Canadian provisions. Nonetheless, one can say that since Canadian provisions are organised in a more detailed way, they are stricter than the U.S. counterpart. In overall, based on the previous comparison, there are not any lessons to be derived for Canada.

As noted previously, CFC rules of the U.S. have been revisited with TCJA. As a consequence of the new GILTI rules, it is possible for Canadian-source income of MNEs to be subject to GILTI tax. This takes place in cases when the tax rate of Canada is below 13,125 per cent due to reductions in the tax rate, e.g., R&D tax credits, which in the outcome

[120] These are corporate income tax rate, group taxation regime, research and development, patent box regime and thin capitalisation rules.
[121] See McKenzie/Smart (2019), pp. 16-17.
[122] See Stratton (2019, p. 19).
[123] See McKenzie/Smart (2019), pp. 16-17.
[124] See Donnelly and Young (2002, pp. 431–432).
[125] See Brauner (2011, pp. 301–301, 311).
[126] See Stratton (2019, p. 11).
[127] See McKenzie and Smart (2019b, pp. 17–19).
[128] See Donnelly and Young (2002, p. 441).
[129] Tax Attractiveness Index (Schanz et al. (2019)) recognises countries that allow loss carryforward for more than 20 years as fully attractive (with the index value "1"). Since Canada allows carryforward up to 20 years, it has the score 0,5 in the index. For the purposes of this thesis, due to the extreme proximity of Canada's allowance to the given threshold, the country can be categorised as attractive.
[130] See The Standing Senate Committee on Banking, Trade and Commerce (2018, pp. 17–18).
[131] See De Luca and Hausch (2017, pp. 59–60).
[132] See The Standing Senate Committee on Banking, Trade and Commerce (2018, pp. 17–18).

lowers the average tax rate on an MNE's worldwide income and makes it subject to GILTI tax. This phenomenon is defined as "treasury transfer effect" since the rate reductions in Canada would be in favour of the U.S. through the GILTI tax.[133] Canadian lawmakers should take this indirect effect into consideration when responding to the U.S. tax reform to avoid any counterproductive consequences.

With the introduction of TCJA, the thin capitalisation rules in the U.S. have been tightened. This can result with U.S. based MNEs shifting their debt to affiliates in other jurisdictions such as Canada, to reduce their debt ratio in the U.S. If the Canadian government does not respond with a policy change, the country would then encounter losses of its tax revenue. Furthermore, it has been also found that tighter thin capitalisation rules result with MNEs substituting regulated forms of debt with unregulated ones. In this alternative strategy, they choose to issue loans in the name of affiliates in jurisdictions with looser rules.[134] Thus, in order to limit aggressive debt shifting towards its territory, Canada should tighten its thin capitalisation rules as well.

As indicated earlier, Canada and the U.S. have, at least on the surface level, very similar transfer pricing rules, whereby both countries' accepted transfer pricing methods are in accordance with the arm's length principle. TCJA's newly introduced provisions such as FDII and GILTI rules had effects on transfer pricing, nevertheless, these effects were indirect. Thus, regarding the transfer pricing rules, there are not any lessons for Canada that can be derived from the provisions of the U.S.

As stated above, with regard to the criteria 3.4.5 "treaty network", Canada has a better standpoint than the U.S., since the country has currently more bilateral tax treaties in force. However, some observers argue that the strong dependence of Canada to its biggest trade partner, the U.S., can be problematic. An example from the recent past is the steel and aluminium tariffs imposed on goods of Canadian exporters by the U.S. government. To eliminate the potential risk that rises due to political turmoil, Canada should try to develop its relationships with other emerging market economies such as China and Mexico.[135]

Apart from the lessons that are derived from the comparison criteria chosen for this thesis, there are other key areas that Canadian lawmakers should consider. Over the years many provisions were added to ITA which ended up with complicating the income tax system. Critics suggest that simplifying ITA would increase the tax attractiveness of Canada in the eyes of the MNEs.[136] When employing these necessary and long overdue changes, the current situation should also be taken into consideration. Since the Coronavirus pandemic is expected to have fundamental changes in societies, the need for an internationally competitive tax system that encourages innovation and investment is more crucial than ever.[137]

5. Conclusion

This thesis aimed to assess the similarities and differences of MNEs' income taxation in the U.S. and Canada after the U.S. tax reform. Based on the 12 corporate taxation criteria that determine the tax attractiveness of jurisdiction, the corporate taxation systems have been compared. It concludes that, as expected, despite many similarities of the fundamental fiscal aspects of both countries, their tax systems differ. It is not possible to uniformly declare the countries as similar or different, since some parameters are very much alike between the two countries, whereas others are completely diverse. These differences have been heightened with the enactment of TCJA. Furthermore, this thesis had the intention to derive lessons for Canada from the U.S. corporate taxation system. By inferring to its own comparison, this objective is fulfilled in a unique and first-of-its-kind manner.

Basic elements differ considerably between both countries, whereby the most significant advantage of Canada, lower corporate income tax rate, disappeared with TCJA. On the other hand, group taxation provision of the U.S. is an important edge over Canada from the perspective of tax attractiveness. With regard to cost recovery, both countries' tax systems were influenced heavily from TCJA, whereby the immediate temporary steps that Canada took with the Fall Economic Update to keep up with the changes of TCJA require further development. The category of tax incentives harboured most of the similarities between the two tax systems. In the future, in order to promote innovation even further in their jurisdictions, the U.S. and Canada should continue to develop their tax incentives. The broad category of international tax regulations demonstrates on the surface level that both countries are strict with their efforts to prevent illegitimate tax planning strategies, but they are also trying to preserve their tax attractiveness for MNEs' investment purposes. With reference to the latter, the new provisions that the U.S. has implemented with the tax reform could be a threat to Canada's tax attractiveness. The last part shows that there are many steps that Canada can take to rebuild its competitive position against the U.S., hence an action of Canadian policymakers in retaliation to TCJA has become necessary.

It is an undeniable fact that the more detailed the conducted analysis is, the larger the divergencies of both countries' corporate taxation systems are going to be; since it is essentially impossible to have two entirely matching taxation systems. But a comparison that considers the fundamental aspects of each criterion allows to see the broader picture. Hence, the methodology chosen for this thesis adequately

[133] See McKenzie and Smart (2019b, pp. 13–15, 20–21).

[134] See Buettner, Overesch, Schreiber, and Wamser (2012, pp. 930–38), as cited in McKenzie and Smart (2019a, p. 8).

[135] See The Standing Senate Committee on Banking, Trade and Commerce (2018, pp. 20–22).

[136] See The Standing Senate Committee on Banking, Trade and Commerce (2018, pp. 9–11).

[137] See Canadian Chamber of Commerce (2020, pp. 1, 8); Canadian Chamber of Commerce (2018, pp. 5–6).

shows similar and different aspects without excessively scrutinising and makes the comparison available for further research.

As stated repeatedly throughout this thesis, based on the conclusions, policymakers of Canada should consider promptly reforming their tax system. Moreover, policymakers worldwide should bear these comparison criteria in mind when trying to assess how attractive their tax systems are from the standpoint of corporate taxation. MNEs can also refer to this comparison when making their own investment decisions. Hence, this comparison between the U.S. and Canada could act as a role model for both governments and businesses. Further research is however needed to actually form the structure of a possible Canadian tax reform. In this regard, before implementing any changes it is also required to assess how the implications of a reform as such are expected to be.

The corporate landscape today is prominent with its great degree of global integration. This makes the previously domestically organised corporate taxation a border-crossing matter. It also obliges countries to arrange their tax systems in a way that it pleases multinational corporations. In this race between governments to make their countries attractive investment hubs, the relativity factor plays a major role, since during decision process companies are often undecided between two jurisdictions. In the future, it is desirable to have tax systems that simultaneously preserve the governments' interests while providing attractive tax environments for companies to invest in. In this regard, the U.S. and Canada have the potential to set the example that lawmakers around the world are hoping to see because as it is impossible to have economic growth without investments, it is also certainly not possible to draw investments without a correct tax policy.

References

Alberta Printed Circuits Ltd. v. The Queen. (2011). *No. 2008- 714(IT)G (Tax Court of Canada)*.

Arnold, B. J. (2012). A Comparative Perspective on the U.S. Controlled Foreign Corporation Rules. *Tax Law Review*, *65*(3), 473–503.

Aronshtam, N., & Kane, M. (2018). *Survey of Global Investment and Innovation Incentives*. Deloitte.

Arrow, K. (1962). *Economic Welfare and the Allocation of Resources for Invention* (Vol. The Rate and Direction of Incentive Activity: Economic and Social Factors; Universities - National Bureau Committee for Economic Research and Committee on Economic Growth of the Social Science Research Council, Ed.).

Atkinson, R. D. (2015). *An Easy Checkoff for Global Competitiveness: The Case for a U.S. Innovation Box*. Information Technology & Innovation Foundation.

Ault, H. J., Arnold, B. J., & Cooper, G. S. (2019). *Comparative Income Taxation: A Structural Analysis* (4th ed.). Alphen aan den Rijn, [ND]: Kluwer Law International.

Avi-Yonah, R. S., & Pichhadze, A. (2017). GAARs and the Nexus between Statutory Interpretation and Legislative Drafting: Lessons for the U.S. from Canada. *Accounting, Economics, and Law*, *7*(1), 165.

Bazel, P., Mintz, J., & Thompson, A. (2018). *2017 Tax Competitiveness Report: The Calm Before the Storm.* doi: http://dx.doi.org/10.11575/sppp.v11i0.43270

Behrens, K., Peralt, S., & Picard, P. M. (2014). Transfer Pricing Rules, OECD Guidelines, and Market Distortions. *Journal of Public Economic Theory*, *16*(4), 650–680. doi: https://doi.org/10.2307/3665873

Berk, J., & DeMarzo, P. (2017). *Corporate Finance* (4th ed.).

Betker, B. L. (1995). An Empirical Examination of Prepackaged Bankruptcy. *Financial Management*, *24*(1), 3–18. doi: https://doi.org/10.2307/3665873

Brackman, S. (2016). *Taxes in the United States : Developments, Analysis and Research*. New York, [US].

Bradley, S., Dauchy, E., & Robinson, L. (2015). Cross-Country Evidence on the Preliminary Effects of Patent Box Regimes on Patent Activity and Ownership. *National Tax Journal*, *68*(4), 1047–1071.

Braun, J., & Fuentes, D. (2016). The Effects of Double Tax Treaties for Developing Countries. A Case Study for Austria's Double Tax Treaty Network. *Public Finance and Management*, *16*(4), 383–433.

Brauner, Y. (2011). Policy Forum: Taxation of Corporate Groups - Lessons from the United States. *Canadian Tax Journal*, *59*(2), 295–314.

Brauner, Y. (2017). Tax Treaty Disputes in the United States. In E. Baistrocchi (Ed.), *A Global Analysis of Tax Treaty Disputes* (pp. 17–64). Cambridge, [UK]. doi: https://doi.org/10.1017/9781316528945.002

Brazell, D. W., & Mackie, J. B. (2000). Depreciation Lives and Methods: Current Issues in the U.S. Capital Cost Recovery System. *National Tax Journal*, *53*(3), 531–561.

Brown, K. B. (2012). Regulation of Corporate Tax Avoidance. In K. B. Brown & D. V. Snyder (Eds.), *General Reports of the XVIIIth Congress of the International Academy of Comparative Law/Rapports Généraux du XVIIIème Congrès de l'Académie Internationale de Droit Comparé* (pp. 609–624). Dordrecht, [NL]. doi: https://doi.org/10.1007/978-94-007-2354-2_27

Buettner, T., Overesch, M., Schreiber, U., & Wamser, G. (2012). The Impact of Thin-Capitalization Rules on the Capital Structure of Multinational Firms. *Journal of Public Economics, Fiscal Federalism*, *96*(11), 930–938. doi: 10.1016/j.jpubeco.2012.06.008

Bunn, D., & Asen, E. (2019). International Tax Competitiveness Index 2019. In *Tax foundation.* Washington D.C. [US].

Burman, L., & Slemrod, J. (2013). *Taxes in America: What Everyone Needs to Know*. New York, [NY].

Canada Revenue Agency. (2005). *Corporation Tax Rates.* Retrieved 2020-05-18, from https://www.canada.ca/en/revenue-agency/services/tax/businesses/topics/corporations/corporation-tax-rates.html

Canada Revenue Agency. (2017). *Line 25200 - Non-Capital Losses of Other Years.* Retrieved 2020-06-16, from https://www.canada.ca/en/revenue-agency/services/tax/individuals/topics/about-your-tax-return/tax-return/completing-a-tax-return/deductions-credits-expenses/line-25200-non-capital-losses-other-years.html

Canada Revenue Agency. (2020). *Scientific Research and Experimental Development Tax Incentive - Overview.* Retrieved 2020-06-08, from https://www.canada.ca/en/revenue-agency/services/scientific-research-experimental-development-tax-incentive-program/overview.html

Canada v. GlaxoSmithKline Inc. (2012). *No. 33874 (Supreme Court of Canada)*.

Canadian Chamber of Commerce. (2018, 10 Ways to Build a Canada That Wins).

Canadian Chamber of Commerce. (2020). *Roadmap to Recovery*.

Christians, A. (2018). *Taxing Transnationals: Canada and the World, Canada in International Law at 150 and Beyond Paper, p. 1–11* (No. 15). Ontario, [CA].

Clark, W. S., & Klemm, A. (2015). Effective Tax Rates for Multinationals - The Role of Tax Incentives and Tax Planning Policy Forum. *Canadian Tax Journal*, *63*(1), 133–148.

Clifford, S. (2019). Taxing Multinationals beyond Borders: Financial and Locational Responses to CFC Rules. *Journal of Public Economics*, *173*, 44–71. doi: 10.1016/j.jpubeco.2019.01.010

Cockfield, A. J. (2017). Tax Treaty Disputes in Canada. In E. Baistrocchi (Ed.), *A Global Analysis of Tax Treaty Disputes* (pp. 116–171). Cambridge, [UK]. doi: https://doi.org/10.1017/9781316528945.004

Correa, P., & Guceri, I. (2013). Tax Incentives for Research and Development, Innovation, Technology and Entrepreneurship Policy Note 4. Washington D.C., [US]: World Bank.

Daniels, J. P., O'Brien, P., & von der Ruhr, M. B. (2015). Bilateral Tax Treaties and US Foreign Direct Investment Financing Modes. *International Tax and Public Finance*, *22*(6), 999–1027. doi: https://doi.org/10.1007/s10797-014-9340-1

Deloitte. (2020). *Jurisdiction: Deloitte International Tax Source.* Retrieved 2020-06-15, from https://www.dits.deloitte.com/#Jurisdiction/98

De Luca, A., Aronshtam, N., Boivin, L., Lamontagne, l.-N., Lassonde, J., McKindsey, D., … Robertson, S. (2016). *Canada's First Patent Box Regime - an Incentive for Domestic R&D Commercialization*. Deloitte, R&D Tax Update.

De Luca, A., & Hausch, J. (2017). Policy Forum: Patent Box Regimes - A Vehicle for Innovation and Sustainable Economic Growth Policy Forum. *Canadian Tax Journal*, *65*(1), 39–60.

Department of Finance Canada. (2018). Investing in Middle Class Jobs.

Department of Finance Canada. (2019). *Tax Treaties.* Retrieved 2020-06-15, from https://www.canada.ca/en/department-finance/programs/tax-policy/tax-treaties.html#status

Donnelly, M., & Young, A. (2002). Policy Options for Tax Loss Treatment: How Does Canada Compare? *Canadian Tax Journal*, *50*(2), 429–488.

Dueñas, S. (2019). CFC Rules Around the World. *Fiscal Fact, Tax Foundation*.

Duke of Westminster v. Commissioners of Inland Revenue. (1935). *No. 19 TC 490 (Court of Appeal England and Wales)*.

Dumalski, M., & DeGrace, D. (2019). Canada Highlights 2019. *Tax Guides and Highlights. Deloitte*.

Ernst & Young LLP. (2019a). *Worldwide Corporate Tax Guide: Canada*.

Ernst & Young LLP. (2019b). *Worldwide Corporate Tax Guide: The United States*.

Fischer, A. (2015). A Comprehensive Approach to Stateless Income Note. *George Washington Law Review*, *83*(3), 1028–1057.

Gale, W., Gelfond, H., Krupkin, A., Mazur, M. J., & Toder, E. (2018). A Preliminary Assessment of the Tax Cuts and Jobs Act of 2017. *National Tax Journal*, *71*(4), 589–612.

Goodlatte, R. B. (2015). *H.R.9 - Innovation Act.*

Goulding, C. R., & Goulding, C. G. (2013). R&D Tax Credit Fundamentals. *Corporate Business Taxation Monthly*, *14*(9), 17–18.

Gregory v. Helvering. (1935). *No. 293 u.s. 465 (Supreme Court of the United States)*.

Harris, P. A., Keen, M., & Li, L. (2019). Policy Forum: International Effects of the 2017 US Tax Reform - A View from the Front Line. Rochester NY, [US]: Social Science Research Network, SSRN Scholarly Paper.

Hashimzade, N., & Epifantseva, Y. (2017). The Routledge Companion to Tax Avoidance Research. *The Routledge Companion to Tax Avoidance Research*. doi: https://doi.org/10.4324/9781315673745

Hoppe, T., Schanz, D., Sturm, S., & Sureth-Sloane, C. (n.d.). *Tax Complexity*

Index. Retrieved 2020-05-20, from https://www.taxcomplexity.org/

Internal Revenue Service. (2019a). *Publication 536 (2019), Net Operating Losses (NOLs) for Individuals, Estates, and Trusts | Internal Revenue Service*. Retrieved 2020-06-16, from https://www.irs.gov/publications/p536

Internal Revenue Service. (2019b). *Publication 542 (01/2019), Corporations | Internal Revenue Service*. Retrieved 2020-05-18, from https://www.irs.gov/publications/p542

Kujinga, B. T. (2012). Analysis of Misuse and Abuse in Terms of the South African General Anti-Avoidance Rule: Lessons from Canada. *The Comparative and International Law Journal of Southern Africa*, *45*(1), 42–63.

Kunka, P., Stasio, A., & Zagler, M. (2020). On the Relevance of Double Tax Treaties. *International Tax and Public Finance*, *27*(3), 575–605. doi: http://dx.doi.org.ezproxy.cul.columbia.edu/10.1007/s10797-019-09570-9

Laplante, S. K., Skaife, H. A., Swenson, L. A., & Wangerin, D. D. (2019). Limits of Tax Regulation: Evidence from Strategic R&D Classification and the R&D Tax Credit. *Journal of Accounting and Public Policy*, *38*(2), 89–105. doi: https://doi.org/10.1016/j.jaccpubpol.2019.02.003

Larsen, P. o., Cordero, L. F., Bowen, M. J., Denault, A. M., & D'Alessandro Jr., P. J. (2018). *Transfer Pricing in the United States: Overview*. Retrieved 2020-06-14, from http://uk.practicallaw.thomsonreuters.com/w-006-9130?transitionType=Default&contextData=(sc.Default)&firstPage=true&bhcp=1

Luscombe, M. A. (2020). Analyzing Net Operating Loss Options in 2020. *The Tax Magazine*, *98*(6), 3–4, 49.

Marley, P., & Tremblay, R. G. (2005). Canada's Offshore Anti-Avoidance Rule Should Receive Narrow Interpretation. *Tax Management International Journal*, *34*(2), 105–108.

McCaffery, E. J. (2012). *The Oxford Introductions to U.S. Law : Income Tax Law*. Oxford, [UK].

McGill, S. P., Karch, G. C., Feeley, K. J., O'Banion E., S., & Crouse, J. G. (2018). GILTI Rules Particularly Onerous for Non-C Corporation CFC Shareholders. *International Tax Journal*, *44*(1), 11–14, 42.

McKenzie, K., & Smart, M. (2019a). Policy Forum: Business Tax Reform in the United States and Canada. *Canadian Tax Journal*, *67*(1), 57–66. doi: 10.32721/ctj.2019.67.1.pf.mckenzie

McKenzie, K., & Smart, M. (2019b). Tax Policy Next to the Elephant: Business Tax Reform in the Wake of the US Tax Cuts and Jobs Act. *C.D. Howe Institute*.

Merlo, V., & Wamser, G. (2014). Debt Shifting and Thin-Capitalisation Rules. *DICE Report*, *12*(4), 27–31.

Merriam-Webster. (n.d.). Definition of Pass-Through Entity.. Retrieved 2020-06-01, from https://www.merriam-webster.com/dictionary/pass-throughentity

Nam, C. W., & Radulescu, D. M. (2005). The Role of Tax Depreciation for Investment Decisions: A Comparison of European Transition Countries. *Eastern European Economics*, *43*(5), 5–24.

Nitikman, J. A. (2000). The Interaction of Canada's Thin Capitalization Rule and the Canada - United States Tax Treaty. *The International Tax Journal*, *26*(1), 20–64.

Nosikova, Y. (2015). *The FAPI Regime and CFC Rules under the BEPS Action Plan 3: Where Does the Business of "Managing Intellectual Property" Stand?* University of Victoria Faculty of Law.

OECD. (2015). *Action 5: Agreement on Modified Nexus Approach for IP Regimes, OECD/G20 Base Erosion and Profit Shifting Project*. Paris, [FR]: OECD Publishing.

OECD. (2017a). *Model Tax Convention on Income and on Capital: Condensed Version 2017*. Paris, [FR]: OECD Publishing. doi: http://dx.doi.org/10.1787/mtc_cond-2017-en

OECD. (2017b). *OECD Transfer Pricing Guidelines for Multinational Enterprises and Tax Administrations*. Paris, [FR]: OECD Publishing. doi: http://dx.doi.org/10.1787/tpg-2017-en

OECD. (2019). *R&D Tax Incentives: United States 2019, Directorate for Science, Technology and Innovation*. Paris, [FR]: OECD Publishing.

OECD. (2020). *Statutory Corporate Income Tax Rate*. Retrieved 2020-06-20, from https://stats.oecd.org/Index.aspx?QueryId=78166

Petrie, J. R. (1952). *The Taxation of Corporate Income in Canada*. Toronto, Ontario [CA].

Pomerleau, K. (2018). *US Corporate Income Tax Now More Competitive*. Retrieved 2020-05-12, from https://taxfoundation.org/us-corporate-income-tax-more-competitive/

Prebble, R., & Prebble, J. (2010). Does the Use of General Anti-Avoidance Rules to Combat Tax Avoidance Breach Principles of the Rule of Law - A Comparative Study Critical Issues in Comparative & International Taxation. *Saint Louis University Law Journal*, *55*(1), 21–46.

PricewaterhouseCoopers International Ltd. (PwC). (2019). *PwC Worldwide Tax Summaries Canada - Corporate - Group Taxation*. Retrieved 2020-06-10, from https://taxsummaries.pwc.com/canada/corporate/group-taxation

PricewaterhouseCoopers International Ltd. (PwC). (2020a). *PwC Worldwide Tax Summaries: United States - Corporate - Deductions*. Retrieved 2020-06-15, from https://taxsummaries.pwc.com/united-states/corporate/deductions

PricewaterhouseCoopers International Ltd. (PwC). (2020b). *PwC Worldwide Tax Summaries United States - Corporate - Group Taxation*. Retrieved 2020-06-07, from https://taxsummaries.pwc.com/united-states/corporate/group-taxation

Sarkar, S. (2014). Valuation of Tax Loss Carryforwards. *Review of Quantitative Finance and Accounting*, *43*(4), 803–828. doi: https://doi.org/10.1007/s11156-013-0393-5

Schanz, D., Keller, S., Dinkel, A., Fritz, J., & Grosselfinger, C. (2017). *Tax Attractiveness Index: Methodology*.

Shafer, J. (2013). Canada Expands Thin Capitalisation Rules. *International Tax Review*, *24*(10), 69–70.

Shaviro, D. (2018). The New Non-Territorial U.S. International Tax System. In *presented at the NYU Law School Oxford Academic Symposium*. New York NY, [US].

Sherlock, M. F., & Marples, D. J. (2019). Overview of the Federal Tax System in 2019. Washington D.C., [US]: Congressional Research Service.

Shoven, J. B., & Whalley, J. (1992). *Canada-U.S. Tax Comparisons, A National Bureau of Economic Research Project Report*. Chicago IL, [US].

Singer, J. M., Bray, R., Kershaw, D., Shulman, M., & Crouch, L. (n.d.). *Claiming Refunds for Corporate Net Operating Loss Carrybacks Under the CARES Act*. Retrieved 2020-06-16, from https://www.shearman.com/perspectives/2020/04/claiming-refunds-for-corporate-net-operating-loss-carrybacks-under-the-cares-act

SRED. (2017). *SR&ED Vs. U.S. R&D Tax Credits: A Comparison*. Retrieved 2020-06-09, from https://www.sreducation.ca/sred-us-rd-comparison/

Stratton, T. (2019). 50 Years of Cutting and Pasting - Modernizing Canada's Tax System. *Canadian Chamber of Commerce*.

Streitferdt, F. G. (2013). The Valuation of Tax Loss Carryforwards. *Journal of Business Valuation and Economic Loss Analysis*, *8*(1), 163–184. doi: http://dx.doi.org.ezproxy.cul.columbia.edu/10.1515/jbvela-2013-0020

Stubart Investments Ltd. v. The Queen. (1984). *No. 16623 (Supreme Court of Canada)*.

Suarez, S. (2019). Transfer Pricing in Canada. *Tax Notes International*, *96*(9), 781–813.

Taklalsingh, R. (2019). *Transfer Pricing Legislation: Effect on Multinational Enterprises in the United States*. Walden University.

Tax Policy Center. (n.d.). *Tax Policy Center's Briefing Book*. Retrieved 2020-06-01, from https://www.taxpolicycenter.org/briefing-book/how-did-tax-cuts-and-jobs-act-change-business-taxes

The Economist. (1993). *"Taxing Questions"* .

The Standing Senate Committee on Banking, Trade and Commerce. (2018). Canada - Still Open for Business? Ottawa ON, [CA].

Thomson Reuters Canada Ltd. (n.d.). *Loss Carry-Over Practical Law Canada*. Retrieved 2020-06-16, from https://ca.practicallaw.thomsonreuters.com/w-021-1214?transitionType=Default&contextData=(sc.Default)

Ting, A. (2013). *The Taxation of Corporate Groups Under Consolidation : An International Comparison*. Cambridge, [UK]: Cambridge University Press.

Traub, J., & Kummer, J. (2015). *Innovation Boxes, International Tax Reform, and Infrastructure Spending. The Pillars of this Fall's Tax Legislative*

Debate. Deloitte.

Tørsløv, T. R., Wier, L. S., & Zucman, G. (2018). *The Missing Profits of Nations*. Cambridge MA, [US]: National Bureau of Economic Research, NBER Working Paper Series w24701.

Trump, C. (2020). *United States Highlights 2020*. Deloitte,Tax Guides and Highlights.

United Nations Conference on Trade and Development (UNCTAD). (2018). *World Investment Report 2018: Investment and New Industrial Policies*. Geneva, [CH]: UNCTAD/WIR/2018.

United States - Canada Income Tax Convention . (1985). Washington D.C. [US].

Waerzeggers, C., & Hillier, C. (2016). *Introducing a General Anti-Avoidance Rule (GAAR) - Ensuring that a GAAR Achieves Its Purpose* (Vol. 1). Tax Law IMF Technical Note, IMF Legal Department.

Wakabayashi, D., & Chen, B. X. (2018). Apple, Capitalizing on New Tax Law, Plans to Bring Billions in Cash Back to U.S. - The New York Times. *The New York Times*. Retrieved 2020-05-26, from https://www.nytimes.com/2018/01/17/technology/apple-tax-bill-repatriate-cash.html

Yang, J. G. S., Lord, R. A., & Saito, Y. (2019). Has the United States Really Changed From a Worldwide to a Territorial Tax System? *Journal of Taxation of Investments*, *36*(4), 69–83.

Yang, J. G. S., Poon, W. W., & Chang, C. (2018). The Important Aspects and Strategies of the TCJA. *Practical Tax Strategies*, *101*(1), 16–23.

Junior Management Science 6(3) (2021) 507-546

Junior Management Science

journal homepage: www.jums.academy

No Mon, No Fun? A Discrete Choice Analysis of the Preferences of Medical Students for Work on the Countryside

Ohne Moos nichts los? Eine Discrete-Choice-Analyse der Präferenzen von Medizinstudenten für die Arbeit auf dem Land

Markus Probst

WHU – Otto Beisheim School of Management

Abstract

The shortage of doctors, especially regarding primary care doctors in rural areas, is one of the greatest challenges the German healthcare system has to face. The aim of this thesis is to investigate which incentives can be used to enable medical students to work in rural areas to ensure medical care there. For this purpose, a discrete choice experiment with a complementary, qualitative survey was conducted. The results indicate (a) a heterogeneity of the medical student population in Germany. Accordingly, some parts of the medical student body would be willing to practice as a country doctor. In addition, (b) it becomes clear that financial incentives to promote this aspiration are only partially effective but can be compensated by other means of support. To illustrate the diversity of decision determinants, an exemplary commitment program is developed to counteract the shortage of rural physicians in Germany.

Zusammenfassung

Der Ärztemangel, insbesondere Hausärzte im ländlichen Bereich betreffend, stellt eine der größten Herausforderungen des deutschen Gesundheitssystems dar. Das Ziel der vorliegenden Arbeit ist es zu untersuchen, welche Anreize für Medizinstudenten geschaffen werden können, um die medizinische Versorgung auf dem Land zu sichern. Hierfür wurde ein Discrete-Choice-Experiment mit einer ergänzenden, qualitativen Umfrage durchgeführt. Die Ergebnisse deuten (a) auf eine Heterogenität der Medizinstudentenschaft in Deutschland hin. Demnach wären einige Teile der Medizinstudentenschaft bereit, als Landarzt zu praktizieren. Zudem wird (b) deutlich, dass finanzielle Anreize zur Förderung dieser Bestrebung nur beschränkt wirksam sind, jedoch durch anderweitige Unterstützung kompensiert werden können. Zur Illustration der Diversität der Entscheidungsdeterminanten wird ein beispielhaftes Verpflichtungsprogramm entwickelt, mit dessen Hilfe dem Landärztemangel in Deutschland entgegengewirkt werden kann.[1]

Keywords: Landärztemangel; Discrete-Choice-Experiment; Medizinstudenten in Deutschland.

1. Einleitung

In unserer schnelllebigen Welt ist unsere Gesundheit eines der wenigen Güter, das unverändert wichtig für uns ist. Gleichwohl ist die Definition des Begriffs Gesundheit nicht unumstritten und wird verschieden interpretiert. Während die meisten Definitionen, wie auch die der Weltgesundheitsorganisation (WHO), sich auf das Wohlergehen eines jeden Menschen fokussiert, wird in anderen Definitionen die Wichtigkeit der Gesundheit für das Funktionieren der Gesellschaft herausgehoben (Parsons, 1967). Abseits ihrer Definition ordnet jedoch auch die WHO Gesundheit in den Kreis der Grundrechte ein und hebt ihren Beitrag zum sozialen Frieden der Gesellschaft hervor (Constitution of the

[1] In dieser Arbeit wird aus Gründen der besseren Lesbarkeit das generische Maskulinum verwendet. Weibliche und anderweitige Geschlechteridentitäten werden dabei ausdrücklich mitgemeint, soweit es für die Aussage erforderlich ist.

Weiterhin wird zur besseren Lesbarkeit von dem „Medizinstudium“ gesprochen. Damit ist lediglich der Studiengang Humanmedizin gemeint. Verwandte Studiengänge wie Veterinärmedizin oder Zahnmedizin sind hierbei nicht eingeschlossen.

DOI: https://doi.org/10.5282/jums/v6i3pp507-546

World Health Organization, 2006).

Um die Gesundheit als Gut aufrechtzuerhalten und zu pflegen, bilden staatliche Gesundheitssysteme einen wichtigen Grundpfeiler. Besonders in Deutschland wird unser weltweit einmaliges System, welches jedem Bürger eine exzellente Versorgung verspricht, häufig herausgestellt. Darüber hinaus verpflichtet das Grundgesetz den deutschen Staat, dem Bürger ein funktionierendes Gesundheitssystem zur Verfügung zu stellen (Wissenschaftlicher Dienst des Deutschen Bundestages, 2015).

Das deutsche Gesundheitssystem bleibt jedoch nicht unberührt von jenen makroökonomischen Entwicklungen, die auch unsere Wirtschaft und Gesellschaft als Ganzes beeinflussen (PricewaterhouseCoopers, 2018). Zunächst abstrakte Konzepte wie die Urbanisierung oder der demografische Wandel wirken sich nicht nur auf Unternehmen aus, sondern auch auf unser Gesundheitssystem. So tritt durch die Kräfte der Urbanisierung eine Verschiebung der Gesellschaft in Richtung Stadt ein (PricewaterhouseCoopers, 2018). Dabei verlassen jedoch nicht alle Gesellschaftsschichten gleichförmig die ländlichen Gebiete Deutschlands, so dass insbesondere Ältere auf dem Land verbleiben (Henger & Oberst, 2019). Ferner wirkt sich auch der demografische Wandel auf das Gesundheitssystem aus, da die Nachfrage nach Gesundheitsleistungen mit dem Alter steigt (Statistisches Bundesamt, 2016). Somit ist von einer steigenden Belastung des Gesundheitssystems und Nachfrage der medizinischen Versorgung, insbesondere auf dem Land, auszugehen. Demgegenüber steht die Angebotsseite der medizinischen Versorgung, welche in Deutschland hauptsächlich durch Ärzte abgedeckt wird. Damit stellt sich der Gesellschaft die Frage, ob aktuell und in Zukunft die Anzahl der Ärzte für die gewünschte Gesundheitsversorgung ausreicht–insbesondere auf dem Land.

1.1. Problemstellung und Relevanz der Arbeit

Diese Frage wird in der Form des Themas „Ärztemangel“ mit verlässlicher Häufigkeit in der öffentlichen Debatte Deutschlands diskutiert. Nach Erhebungen und Vergleichen der Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (OECD) muss dieser Begriff jedoch relativiert werden (Ono, Schoenstein, & Buchan, 2014). Innerhalb der entwickelten Länder liegt Deutschland mit einer Ärztedichte von 4 Ärzten pro 1000 Einwohner im internationalen Vergleich weit vorne (OECD, 2016). Der Begriff „Ärztemangel“ manifestiert sich jedoch in einer regionalen Ungleichverteilung zwischen ländlichen und städtischen Gebieten (Günther, Kürstein, Riedel-Heller, & König, 2010; Ono et al., 2014). Auf dem Land ist die Zahl der Allgemeinärzte und Spezialisten deutlich geringer als in der Stadt (z.B. Roick et al., 2012), so dass häufig von einem „Landärztemangel“ gesprochen wird. Somit verändert sich auch die Natur des Problems von einem absoluten Unterangebot zu einem Verteilungsproblem. Damit ist Deutschland weltweit gesehen nicht alleine. Andere OECD-Länder wie die USA, das Vereinigte Königreich, Kanada und Australien (z.B. Günther et al., 2010; Joyce, McNeil, & Stoelwinder, 2004; Robertson, Boyd, Hedges, & Keenan, 2007; Rosenthal, Zaslavsky, & Newhouse, 2005) zeigen ähnliche Phänomene. Das Gleiche gilt für einige Schwellen- und Entwicklungsländer, wie Thailand, Indonesien sowie viele Länder Afrikas und Lateinamerikas, auch wenn die medizinische Ausgangssituation in jenen Nationen eine andere ist (Dussault & Franceschini, 2006). Somit lässt sich der Landärztemangel als Problem von globaler Tragweite klassifizieren (Starfield, Shi, & Macinko, 2005).

Die Auswirkungen des Landärztemangels sind mannigfaltig und hängen stark von der Schwere des Bedarfs und der Struktur des Gesundheitssystem ab. Jedoch kann generell festgestellt werden, dass die Gesundheit der Bewohner ländlicher Gebiete durch den Mangel beeinträchtigt wird, was zu einer verkürzten Lebenserwartung, verringerten Lebensqualität und hohen Folgekosten für das Gesundheitssystem führt (Guanais & Macinko, 2009). Insbesondere Letztes betrifft die Gesellschaft als Ganzes. Ferner wird dadurch die Ungleichheit in der Bevölkerung im Gefälle zwischen Stadt und Land verstärkt.

Die Literatur beschreibt eine Vielzahl an Gründen, welche für Deutschland z.T. im Ausbildungs- und Gesundheitssystem gesehen werden (Holst, Normann, & Herrmann, 2015; Kasch et al., 2016; Roick et al., 2012). Teilweise der Komplexität der Thematik geschuldet, zeichnet sich nichtsdestotrotz nur ein Bild mit wenig Zusammenhängen und noch weniger Lösungsansätzen. Genauso wenig wird der zeitliche Aspekt berücksichtigt, so dass, entgegen der Empfehlungen der OECD, zum Landärztemangel weniger die zukünftigen Ärzte als die gegenwärtigen Ärzte in den Vordergrund gestellt werden (Ono et al., 2014).

Daraus ergibt sich die folgende Forschungsfrage, die in dieser Arbeit betrachtet werden soll:

> In wieweit können Anreizsysteme für Medizinstudenten strukturiert werden, damit diese nach ihrer Approbation in ländlichen Gebieten arbeiten, um dem dort vorherrschenden Ärztemangel entgegenzuwirken?

1.2. Zielsetzung der Arbeit

Diese Arbeit hat das Hauptanliegen, das Problem des „Landärztemangels“ lösungsorientiert zu untersuchen. Die bisherige Literatur hat in verschiedenen Studien weltweit den Landärztemangel beschrieben und untersucht. Der Fokus liegt dort auf existierenden Landärzten und Korrelationen der Ärztedichte mit Gebietsmerkmalen oder Merkmalen der Ärzte. Dies vernachlässigt nicht nur die zukünftige Generation an Ärzten, die heutigen Medizinstudenten, sondern auch deren Präferenzen bezüglich ihrer Arbeitsstellen. Um die Forschungsfrage in den Kontext der bisherigen wissenschaftlichen Forschung einzuordnen, wird zu Beginn dieser Arbeit ein Überblick über die bestehende Literatur gegeben.

Daraufhin wird für die Situation in Deutschland ein Discrete-Choice-Experiment (DCE) mit Medizinstudenten

durchgeführt, bei dem den Probanden wiederholt Angebote gemacht werden, sich für eine monatliche Zahlung während des Studiums für die Arbeit als Landarzt zu verpflichten. Somit werden mögliche Jobangebote simuliert und Präferenzen bezüglich der Arbeit auf dem Land können herausgelesen werden. Aufgrund ihrer Realitätsnähe stellen DCEs ein bevorzugtes Instrument in der Forschung der Gesundheitsökonomie dar (Ryan, Gerard, & Amaya-Amaya, 2008).

Weiterhin werden sogenannte Willingness-to-Accept Werte (WTA) geschätzt. Diese geben an, gegen welche Zahlung ein Student sich für eines der Angebote verpflichten würde. Verwandt dazu ist die Berechnung der Wahrscheinlichkeit, dass ein Angebot angenommen wird. Durch die richtige Zusammenstellung eines Angebots können somit bereits existierende Programme zur Verpflichtung als Landarzt untersucht und ihre Erfolgswahrscheinlichkeit abgeschätzt werden.

Danach werden weitere Ergebnisse präsentiert und in den Rahmen der bestehenden Forschung eingeordnet. Hierbei liegt der Fokus speziell auf der Literatur bezüglich der Einflussfaktoren auf die Rekrutierung und Erhaltung von Landärzten. Weiterhin wird im Rahmen der Handlungsempfehlungen ein deutschlandweites Verpflichtungsprogramm skizziert, welches an den deutschen Staat und weitere Institutionen, wie Kassenärztlichen Vereinigungen und Ärztekammern, adressiert ist. Dieses soll es ermöglich, deutschlandweit die medizinische Grundversorgung zu verbessern sowie ein besseres Verständnis der Interessen und Bedürfnisse der kommenden Arztgeneration zu erreichen.

Anschließend werden Limitationen diskutiert und ein Ausblick für die weitere Forschung gegeben, bevor diese Arbeit abschließend zusammengefasst wird.

1.3. Hintergrund

Zu dem Thema des Landärztemangels soll in diesem Unterkapitel eine kurze Einführung gegeben werden. Der Landärztemangel und die Frage, wie dieser behoben werden kann, beschäftigen Politik und Gesellschaft in Deutschland schon seit geraumer Zeit. Daher weist die ärztlich-medizinische Landschaft eine Fülle an Ideen und Instrumenten auf, die mit mehr oder weniger Erfolg angewendet wurden.

Zuallererst gilt es zu verstehen, was einen Landarzt ausmacht und was diesen von anderen Ärzten unterscheidet. Während der Begriff häufig emotional konnotiert ist, sei es romantisiert wie in Kafkas gleichnamiger Kurzgeschichte oder in letzter Zeit auch z.T. abwertend, existiert keine einheitliche und verlässliche Definition. Während in der wissenschaftlichen Literatur keine oder nur übermäßig technische Definitionen angewendet werden (Marschner, Wasem, & Heymann, 2018), wird in der öffentlichen Debatte der Begriff des Landarztes lediglich implizit definiert (KBV, 2018b). In Anlehnung an Beschreibungen der Kassenärztlichen Bundesvereinigung (KBV) soll für diese Arbeit ein Landarzt als Folgender definiert werden: Ein Landarzt ist entweder ein Allgemeinmediziner oder Facharzt (z.B. Internist oder Kinderarzt), welcher in seiner Funktion als Hausarzt die medizinische Grundversorgung (Primärversorgung) in ländlichen Gebieten sicherstellt. Dieser hat vielfältige Aufgaben. Er stellt den ersten Anlaufpunkt für medizinische Fragen jeglicher Art dar, unterhält langfristige und persönliche Beziehungen mit seinen Patienten, versorgt grundlegende medizinische Bedürfnisse auf gesamtheitliche Art und Weise und koordiniert die Arbeit von Spezialisten, wenn benötigt (Starfield et al., 2005, p. 458). Der Begriff des Landarztes soll für diese Arbeit trotz oben genannter emotionaler Konnotationen beibehalten werden, um die Einzigartigkeit seiner Arbeit und seiner Integration in die ländliche Gemeinschaft herausstellen. Damit ist ein Landarzt mehr als „nur“ ein Arzt, der in ländlichen Gebieten praktiziert.

Über die Existenz eines Landärztemangels wurde häufig diskutiert und seine Existenz in Deutschland wurde mehrfach belegt. Aufgrund der Komplexität des Themas ist es jedoch schwierig, ein eindeutiges Bild zu zeichnen. Verschiedene Institutionen von Ministerien bis zu Ärztekammern und Kassenärztlichen Vereinigungen erfassen Daten nach unterschiedlichen Kriterien. Das deutsche Krankenversicherungssystem sorgt zusätzlich dafür, dass nicht alle Ärzte für alle Bürger gleich zugänglich sind (Sundmacher & Ozegowski, 2016). Daher existiert eine Vielzahl an Studien, die alle leicht verschiedene Ergebnisse zeigen. Einige Studien (z.B. Roick et al., 2012) nutzen für den Nachweis die Zahlen der Bundesärztekammer und die Metrik des Versorgungsgrades, welcher den prozentualen Unterschied zwischen Soll- und Ist-Zustand der Bedarfsplanung der Kassenärztlichen Verbände beschreibt. Dabei wird bei dem Soll-Zustand von der benötigten medizinischen Grundversorgung ausgegangen. In Deutschland ist der Versorgungsgrad der Allgemeinmediziner als Hausärzte im Durchschnitt geringer als der Versorgungsgrad von anderen Fachärzten. Die KBV weist hier für 289 von 971 von ihr definierten Mittelregionen in Deutschland eine hausärztlichen Versorgungsgrad von unter 100%, d.h. eine Unterversorgung, aus (KBV, 2018b). Deutlich wird auch die regionale Diskrepanz zwischen Nord- und Süddeutschland, wobei Süddeutschland eine wesentlich bessere Versorgungslage hat (Roick et al., 2012). In den neuen Bundesländern ist der Versorgungsgrad häufig geringer als in den alten Bundesländern (Kistemann & Schröer, 2007). Auch wenn diese Metrik nicht alle Aspekte eines Ärztemangels bzw. Landärztemangels abdeckt, kann sie als guter Ausgangspunkt für weitere Analysen genutzt werden. Andere Studien analysieren die Ärztedichte, d.h. Ärzte pro Einwohner, und zeigen dort ähnliche Defizite auf (Sundmacher & Ozegowski, 2016). Auch wenn sich ein Landärztemangel nicht vollkommen pauschalieren und eindeutig messen lässt, zeigt sich jedoch deutlich ein Trend hin zu einem Mangel an Hausärzten auf dem Land; kurz ein Landärztemangel ist in Deutschland vorhanden.

Seit 2019 existiert die sogenannte Landarztquote als Instrument, um dem Landärztemangel aktiv entgegenzuwirken. Dabei wird hier in den Auswahl- und Zulassungsprozess für das Medizinstudium in Deutschland eingegriffen, indem eine bestimmte, von den Bundesländern abhängige Prozentzahl der Studienplätze für zukünftige

Landärzte reserviert wird. Die Studenten verpflichten sich vertraglich im Gegenzug für einen Zeitraum, der vom Bundesland abhängig ist, aber i.d.R. mindestens 10 Jahre beträgt, in ländlichen bzw. ärztlich unterversorgten Gebieten zu arbeiten (Hartmannbund, 2019). Sollte der Arzt seine Verpflichtung nicht erfüllen, gibt es die Möglichkeit, gegen eine Zahlung einer hohen Vertragsstrafe, z.B. in Rheinland-Pfalz bis zu 250000 € (Landesgesetz zur Sicherstellung der hausärztlichen Versorgung in ländlichen Regionen des Landes Rheinland-Pfalz (LArztG RP), 2019), den Vertrag aufzulösen. Diese Landarztquote ist nach einer ähnlichen Quote, durch welche die Bundeswehr Ärzte für ihren Dienst verpflichtet, gestaltet. Nach der Einführung in Nordrhein-Westfalen haben auch weitere Bundesländer wie Baden-Württemberg, Bayern, Rheinland-Pfalz, das Saarland sowie Sachsen-Anhalt eine sogenannte Landarztquote beschlossen. Andere Bundesländer, darunter Hessen, Niedersachsen und Sachsen, haben noch keine Beschlüsse gefasst, untersuchen aber die Sachlage aktuell (Hartmannbund, 2020). Da Hochschulbildung und damit auch die Zulassung zum Studium zum großen Teil in den Händen der Länder liegt, unterscheiden sich die Programme.

In der öffentlichen Diskussion hat die Einführung der Landarztquote gemischte Reaktionen hervorgerufen. Befürworter sehen sie als gute Möglichkeit, um dem Landärztemangel entgegenzuwirken. Zusätzlichen nehmen sie die Landarztquote als Chance für diversere Medizinstudenten wahr, da in Auswahlgesprächen zur Zulassung zum Studium andere Kriterien neben der Abiturnote Berücksichtigung finden (Medical Tribune, 2019). Kritiker weisen auf den langen Zeitrahmen der Verpflichtung von 10 Jahren hin, welcher, relativ zum Lebensalter der Verpflichteten, hoch sei (Hartmannbund, 2019). Die Konsequenz davon seien Fehlanreize und in vielen Fällen unglückliche und unmotivierte Landärzte (Hartmannbund, 2019). Weiterhin befürchten sie eine weitere Verkomplizierung des Zulassungsprozesses und warnen vor der Möglichkeit, dass Medizinstudienplätze gekauft werden können, indem die Vertragsstrafe in Kauf genommen wird (Bauer, 2017). Für eine abschließende Beurteilung ist die Landarztquote jedoch noch zu neu. In keinem Bundesland haben Studenten, die ihren Studienplatz über diese Quote erhalten haben, das Studium abgeschlossen. Somit praktiziert zum jetzigen Zeitpunkt noch kein „Quotenlandarzt“.

Weiterhin wurden auch Verpflichtungsprogramme für zukünftige Landärzte eingeführt, welche in der Öffentlichkeit deutlich weniger Beachtung gefunden haben. Diese werden häufig als „Stipendium“ beschrieben und beinhalten eine monatliche Zahlung während des Studiums dafür, dass sich der Student verpflichtet, nach dem Ende des Studiums auf dem Land zu arbeiten. In Bayern, Brandenburg, Mecklenburg-Vorpommern und Niedersachsen erhalten Studenten für diese Form der Verpflichtung einen monatlichen Betrag zwischen 300 € und 1000 € . Im Falle der Auflösung der Verpflichtung müssen die erhaltenen Leistungen zurückgezahlt werden (Bauer, 2017). Das Programm in Brandenburg wird im Verlauf dieser Arbeit exemplarisch verwendet, um mit den Ergebnissen der Umfrage quantitativ analysiert zu werden. Dabei werden die durchschnittlich benötigte Zahlung für die Teilnahme am Programm (die sogenannte Willingness-to-Accept) sowie die Akzeptanzwahrscheinlichkeiten für das bestehende Programm berechnet. Andere Bundesländer und die zugehörigen Kassenärztlichen Vereinigungen (KVen) setzen ihre finanziellen Mittel an späteren Stellen, wie z.B. der Facharztweiterbildung oder Niederlassung ein. Teilweise findet die Förderung auch auf anderer Ebene statt. So hat der hessische Landkreis Waldeck-Frankenberg die Initiative „Landarzt werden“ gestartet und wirbt gezielt für den Landkreis Ärzte an; auch unter Zuhilfenahme eines Verpflichtungsprogramms (Landkreis Waldeck-Frankenberg, 2017). Die folgende Abbildung 1 zeigt eine Übersicht über existierende Landarztquoten und -verpflichtungsprogramme während des Studiums. Dabei werden aufgrund der fragmentierten Datenlage nur Programme auf Ebene der Länder abgebildet.

Es wird deutlich sichtbar, dass die Landarztquote als Instrument weitverbreitet ist. Dies korreliert mit dem Stand der öffentlichen Diskussion, bei der Verpflichtungsprogramme kaum angesprochen werden. Im Gegensatz zur Landarztquote stellen sie einen geringeren Eingriff in die Gesundheits- und Bildungspolitik dar, ändern jedoch die Zusammensetzung und Eigenschaften der Medizinstudentenschaft nicht. Weiterhin müssen aufgrund ihrer Natur mehr finanzielle Mittel aufgebracht werden. Bei der Betrachtung der Konsequenzen eines Landärztemangels sollten diese jedoch relativiert werden und zu rechtfertigen sein.

In Deutschland ist der Großteil dieser Programme und Quoten jedoch noch neu, so dass auch hier die Beurteilung des Erfolgs zum jetzigen Zeitpunkt noch aussteht.

2. Literaturüberblick

Dieser erste Teil der vorliegenden Arbeit setzt sich mit der bestehenden Literatur auseinander. Anknüpfend an die vorhergehende Einleitung werden tiefergehende Konzepte beschrieben und verschiedene Positionen zu den relevanten Thematiken aufgezeigt. Damit soll eine Basis für ein gemeinsames Verständnis aufgebaut werden, welches für den weiteren Verlauf der Arbeit, und insbesondere der Diskussion unabdinglich ist.

Die Existenz eines Landärztemangels wurde bereits in der Einleitung anhand der Zahlen der KBV und die darauf aufbauenden Studien kurz belegt und soll nicht den Kern dieser Arbeit darstellen. Daher werden nun die mannigfaltigen Gründe für einen Landärztemangel beschrieben. Dabei wird der Fokus auf das Zusammenspiel der Vielzahl von Faktoren, die einen Arzt dazu bringen, als Landarzt zu praktizieren, gelegt. Danach werden die Auswirkungen des Landärztemangels auf das Gesundheitssystem und die Bürger veranschaulicht. Zum Abschluss werden Systeme und Programme andere Länder, mit denen jene versuchen, einen Landärztemangel zu bekämpfen, aufgeführt.

Die untersuchte Literatur zu diesen Themen ist aufgrund der globalen Tragweite breit gefächert. Wichtig für

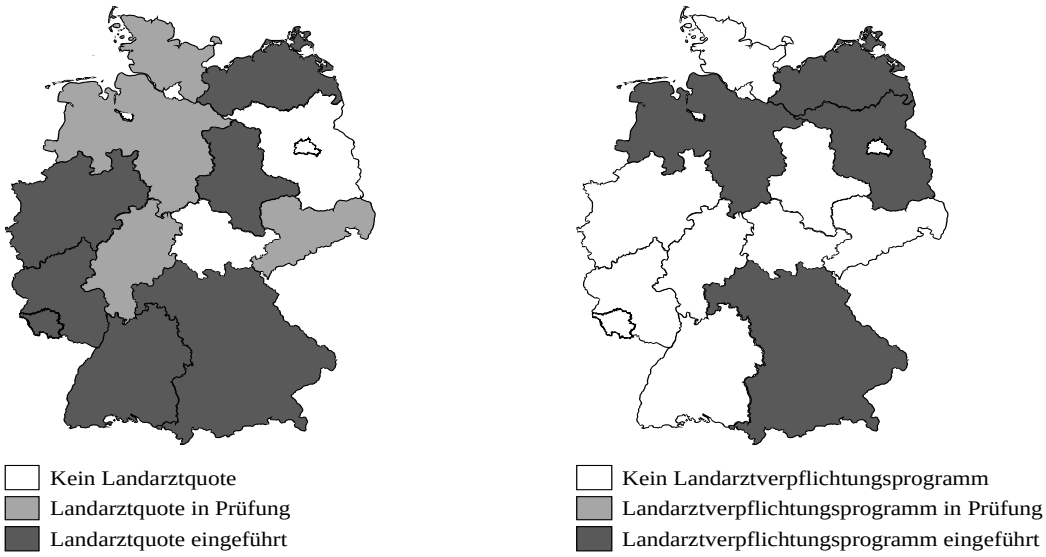


Figure 1: Übersicht über bestehende Landarztförderungen während des Studiums

Quelle: Bauer (2017), Hartmannbund (2020), KV Niedersachsen (2017)

die Analyse und Vergleichbarkeit internationaler Studien ist auch die Unterscheidung in den Aufgaben von Ärzten und Krankenpflegern (engl. nurses). Die Aufgaben und Kompetenzen solcher Krankenpfleger decken, besonders in angelsächsischen Ländern, auch die Diagnose, Überweisungen von Patienten oder Verschreibung von Medikamenten ab (Maier & Aiken, 2016). Diese Aufgaben und Kompetenzen sind, z.T. mit jenen von Landärzten kongruent. Daher können Ergebnisse aus Studien zu Krankenpflegern in ländlichen Regionen, mit Vorbehalten, auch betrachtet werden.

2.1. Einflussfaktoren auf den Landärztemangel

Die Gründe für den Mangel an Landärzten lassen sich anhand des Werdegangs eines diesen in mehrere Phasen unterteilen. Während einige Studien von zwei Phasen reden (Pathman, Konrad, & Agnew, 1994), sprechen andere Studien von drei, wobei häufig die erste Phase weiter aufgeteilt wird (Dussault & Franceschini, 2006). Im ersten Teil der ersten Phase müssen Mediziner Interesse daran zeigen, Landarzt zu werden, beziehungsweise dieses aufbauen („Attraktion“). Danach folgt im zweiten Teil die Einrichtung einer Praxis auf dem Land („Installation“). Viele Studien betrachten diese erste Phase zusammenfassend als „Rekrutierung“. Über die darauffolgende Zeit gilt es den Mediziner als Landarzt zu halten, im Englischen „retention“, damit dieser weiterhin dort praktiziert („Erhaltung“). Die zwei Phasen haben teilweise unterschiedliche Gründe und Einflussfaktoren (Pathman et al., 1994). Wichtig hierbei ist jedoch das Zusammenspiel der beiden Effekte. Ein Mangel an Landärzten kann so durch zu wenig Rekrutierung oder zu geringe Erhaltung hervorgerufen werden (Pathman et al., 1994). Für Deutschland zeigt sich die Rekrutierung als besonderer Schwerpunkt (Roick et al., 2012), um das Problem eines Landärztemangels anzugehen. Neben der schwierigen Rekrutierung wird die Situation in Deutschland durch das hohe Durchschnittsalter der Landärzte gefährdet. Viele der Landärzte stehen kurz vor der Pensionierung und senken damit altersbedingt die Erhaltungsquote (Roick et al., 2012).

Demzufolge werden im Folgenden verschiedene Gruppen von Faktoren unterschieden und einzeln vorgestellt. Die Charakterisierung folgt Dussault and Franceschini (2006) und ordnet weitere Studien in ihr Schema ein. Beginnend bei individuellen Faktoren, werden danach organisatorische und professionelle Faktoren analysiert. Anschließend werden Faktoren bedingt durch die Ausbildungs-, Gesundheits- und andere Systeme und deren Institutionen beleuchtet. Zuletzt wird das soziokulturelle Umfeld mit seinen auf den Landarzt einwirkenden Faktoren betrachtet. Wo möglich, wird eine Unterscheidung in die obengenannten Phasen der Rekrutierung und Erhaltung getroffen.

Individuelle Faktoren

Dussault and Franceschini (2006) beschreiben individuelle Faktoren, welche die Entscheidung auf dem Land zu praktizieren, beeinflussen. Neben demografischen Kriterien wie dem Geschlecht, sind auch soziale Herkunft und die Rahmenbedingungen der medizinischen Ausbildung relevant. Viele dieser Faktoren zeigen wechselseitige Abhängigkeiten untereinander, was im Rahmen des Entwicklungsmodells von Hancock, Steinbach, Nesbitt, Adler, and Auerswald (2009) diskutiert wird.

Jüngere Ärzte haben eine höhere Wahrscheinlichkeit, zu Beginn ihres Berufslebens auf dem Land zu arbeiten. Dieser Unterschied liegt in der geringeren Anzahl an Abhängigkeiten vom Partner, Kindern oder anderen Personen, die diese haben (Kolstad, 2011). Somit fällt es ihnen leichter, Arbeitsort und -stätte zu wechseln, was der Rekru-

tierungsquote zu Gute kommt. Im Gegensatz dazu kann dieser altersbedingte Effekt sich jedoch nachteilig auf die Erhaltung auswirken, da Präferenzen sich ändern und neue Abhängigkeiten entstehen können (Roick et al., 2012).

Bei der weltweiten Betrachtung zeigt sich ein geringerer Anteil an weiblichen Landärzten als an ihren männlichen Kollegen. So beschreiben Doescher, Ellsbury, and Hart (2000) für die Situation in den USA, Szafran, Crutcher, and Chaytors (2001) für Kanada und Matsumoto, Okayama, Inoue, and Kajii (2005) für Japan einen signifikanten Unterschied. Als Hauptgrund für diese Diskrepanz wird der höhere Fokus auf die Bedürfnisse des Partners, welcher unter Umständen nicht in einem ländlichen Gebieten leben und arbeiten möchte, genannt (Szafran et al., 2001). Für männliche Ärzte hingegen werden die Bedürfnisse des Partners bzw. der Partnerin in diesen Studien als weniger wichtig eingestuft. Außerdem zeigen weibliche Ärzte eine höhere Präferenz für ein festes Gehalt, das schwerer mit dem Praktizieren als Landarzt vereinbar ist (Dussault & Franceschini, 2006). Neuere Zahlen aus Deutschland zeigen jedoch hierzulande ein anderes Bild. Hier besteht eine leichte Tendenz zu weiblichen Landärzten (Roick et al., 2012). Ob der Unterschied geographischer Natur (Deutschland vs. USA, bzw. Kanada) ist und durch soziokulturelle Unterschiede, wie die Rolle der Frau, erklärt werden kann, oder doch zeitlicher Natur ist (Studien von 2012 vs. 2000, bzw. 2001), lässt sich nicht feststellen.

Die Bedürfnisse des Partners und der Familie sind, wie vorhergehend umrissen, ein weiterer wichtiger Einflussfaktor. Insbesondere sind der geeignete Arbeitsplatz für den Partner und, sofern Kinder geplant oder vorhanden sind, angemessene Bildungs- und Betreuungseinrichtungen wichtig für die Ortswahl. Gebiete, in denen Landärzte praktizieren, weisen häufig weniger Jobmöglichkeiten und Bildungseinrichtungen auf Marschner et al., 2018. Eine Vielzahl an Studien sieht diesen Mangel als ausschlaggebend für die Rekrutierung und Erhaltung von Landärzten (z.B. Daniels, Van Leit, Skipper, Sanders, & Rhyne, 2007; Han & Humphreys, 2006; Mayo & Mathews, 2006; Szafran et al., 2001). Dieser Effekt konnte auch für Deutschland nachgewiesen werden (Roick et al., 2012).

Eine „ländliche Herkunft“ ist einer der Faktoren, der die Wahrscheinlichkeit auf dem Land zu praktizieren, am stärksten positiv beeinflusst. Dabei bietet der Begriff der ländlichen Herkunft einen breiten Interpretationsspielraum, der ebenfalls im Detail untersucht wird. So existieren Studien, die den Geburtsort praktizierender Landärzte untersuchen (z.B. Geyman, Hart, Norris, & Lishner, 2000; Laven & Wilkinson, 2003). Andere betrachten den Ort des Aufwachsens, Urlaubsorte während der Kindheit oder andere ländliche Kindheitserfahrungen (z.B. Hancock et al., 2009). Der aktuelle Wohnort der Familie wird untersucht und in den Begriff der ländlichen Herkunft miteinbezogen (z.B. Daniels et al., 2007; Szafran et al., 2001). Wieder andere Studien fokussieren sich auf den Ort des Medizinstudiums oder der damit verbundenen Praktika (z.B. Daniels et al., 2007; Herd, Bulsara, Jones, & Mak, 2017). Allen Studien ist jedoch gemeinsam, dass sie „ländliche Herkunft“ als einen besonderen Bezug zum Land sehen. Durch diesen Bezug entsteht eine positive Beziehung zwischen „ländlicher Herkunft“ und der Rekrutierung von Landärzten (Hancock et al., 2009). Dies tritt unabhängig vom Land auf. So zeigt sich der gleiche Effekt für die USA, Kanada, Japan (Matsumoto et al., 2005) als auch für Deutschland (Roick et al., 2012), wobei jedoch die Ausprägung schwer verglichen werden kann.

Diese Vielzahl an Möglichkeiten einer ländlichen Herkunft bzw. der Bezug zum Leben auf dem Land lässt sich im Entwicklungsmodell von Hancock et al. (2009) einordnen und beschreiben. Sie beschreiben vier Entwicklungspfade innerhalb dieses Modells; wobei sich die Pfade der (1) Familiarität, (2) Gemeinschaft und (3) Selbstverwirklichung zu den individuellen Faktoren rechnen lassen. Der vierte Pfad wird später aufgegriffen. Jeder dieser Pfade beinhaltet Faktoren für die Rekrutierung und Erhaltung von Landärzten und ist verflochten mit der ländlichen Herkunft und dem Bezug zum Land.

Beim ersten Pfad, der „Familiarität“, schätzen Landärzte besonders die Gefühle von Vertrauen, Komfort und Gewohnheit, welche ihre Tätigkeit auf dem Land bietet. Diese führen schlussendlich auch zu einer höheren Erhaltungsquote als Landarzt (Hancock et al., 2009).

Der zweite Pfad lässt sich unter dem Begriff „Gemeinschaft“ zusammenfassen. Der Wunsch, etwas für die Gemeinschaft zu tun oder langfristige und tiefergehende Arzt-Patienten-Beziehungen zu pflegen, ist hierbei die treibende Kraft. Ärzte, die diesem Pfad folgen, zeichnen sich durch eine aktive soziale Beteiligung in der ländlichen Gemeinschaft aus.

Der dritte Pfad ist die „Selbstverwirklichung“. Dieser baut auf der Theorie der Bedürfnishierarchie auf, bei der Selbstverwirklichung die höchste Bedürfnisstufe darstellt (Maslow, 1954). Vorherige Bedürfnisstufen, welche von physiologischen Bedürfnissen wie Hunger über Sicherheits-, soziale und Individualbedürfnisse reichen, müssen vorher befriedigt werden (Maslow, 1954). Trotz der Kritik an Maslows Arbeit zeigte sich dieser Pfad im Modell von Hancock et al. (2009) als relevant. Andere Studien (z.B. Szafran et al., 2001) haben ähnliche Ergebnisse und sehen Selbstverwirklichung, insbesondere bei männlichen Ärzten, als relevant an.

In ihrer Studie sehen Hancock et al. (2009) jeden der Pfade als etwa gleich häufig erwähnt und weisen auf mögliche Überlappungen hin. Als Kennzeichen eines positiv durchlaufenen Pfades werden Verwurzelung (in der Ländlichkeit), Wohlbefinden, Zufriedenheit, Verbundenheit sowie besondere Aspekte der Eigenidentität beschrieben (Hancock et al., 2009). Die Komplexität und Vielfältigkeit des Entwicklungsmodells veranschaulichen die Heterogenität des Landarztdaseins.

Organisatorische und professionelle Faktoren

Die nächste Kategorie von Einflussfaktoren sind organisatorische und professionelle Gründe (Dussault & Franceschini, 2006). Diese umfassen Aspekte wie Gehälter und

Gehaltsabstufungen, sowie die Art der Arbeit selbst, aber auch Möglichkeiten zur Weiterbildung.

Die Vergütung eines Landarztes und der Effekt auf die Rekrutierung und Erhaltung von Landärzten wird dabei in vielen Studien untersucht, wobei die Ergebnisse ein gemischtes Bild zeichnen. Generell beschreibt die Mehrzahl der Studien eine geringe, aber trotzdem positive Korrelation zwischen Vergütung und der Arbeit auf dem Land (Dussault & Franceschini, 2006; Fruen, Hadley, & Korper, 1980; Kristiansen & Førde, 1992). Eine höhere Vergütung für die Arbeit als Landarzt ist somit mit einer höheren Rekrutierungs- und Erhaltungsquote assoziiert. Aus der Literatur quantitative Schlüsse auf die Höhe der Vergütung zu ziehen, erweist sich als schwierig, da Zeitpunkt, Länder, Testgruppen und betrachtete Vergütungsänderungen sehr heterogen sind. Das nominale Einkommen spielt für die Rekrutierung von Landärzten eine Rolle, da hierbei aktiv verschiedene Stellen oder Niederlassungsmöglichkeiten miteinander verglichen werden (Kasch et al., 2016). Bei der Erhaltung von Landärzten kommt es jedoch viel mehr auf die Vergütung relativ zu den Lebenshaltungskosten an. Da diese auf dem Land tendenziell niedriger sind, kann hier auch das Einkommen zu einem bestimmten Grad niedriger sein, ohne die Erhaltung negativ zu beeinflussen (Fruen et al., 1980). Jedoch wird laut van den Bussche, Schwartz, and Angerer (2010) in einem deutschlandweiten Vergleich das als zu gering wahrgenommene Einkommen von Landärzten als negativer Aspekt kritisiert. Generell wird das Einkommen als relevanterer Aspekt sowohl für Männer als auch für Frauen beschrieben (Szafran et al., 2001).

Eine zu hohe Arbeitszeit ist ein weiterer Faktor, der sich negativ auf die Erhaltung auswirken kann. In seiner Eigenschaft als Zufriedenheitsfaktor wird in einigen Studien ein hoher Arbeitsaufwand als negativer Einfluss auf die Erhaltung von Landärzten beschrieben (Daniels et al., 2007; Hegney, Mccarthy, Rogers-Clark, & Gorman, 2002). Auch kann er in der Rekrutierung abschreckend wirken und damit auf diese negativ einwirken (Kasch et al., 2016). Weiterhin ist die Zahl der gearbeiteten Stunden auch für die Zufriedenheit des Partners oder der Partnerin wichtig (Mayo & Mathews, 2006). Dies kann, wie im Teil über individuelle Faktoren beschrieben, auch Auswirkungen auf die Erhaltung von Landärzten haben. Für weibliche Ärzte wird dieser Effekt als stärker als für ihre männlichen Kollegen beschrieben (Szafran et al., 2001).

Neben der reinen Arbeitszeit können sich auch Bereitschaftsdienste negativ auf die Rekrutierung und Erhaltung von Landärzten auswirken. In bestehenden DCEs (z.B. Scott et al., 2013) wurden neben der Arbeitszeit auch Bereitschaftsdienste als Eigenschaft der Stelle inkludiert. Dabei wurde ersichtlich, dass sowohl die Zahl der Bereitschaftsdienste als auch die Häufigkeit der Anrufe wichtige Faktoren der Zufriedenheit eines Arztes mit seiner Stelle darstellen (Ubach, Scott, French, Awramenko, & Needham, 2003). Dies beeinflusst die Erhaltung von Landärzten und, durch eine negativere Wahrnehmung der Stelle, auch die Rekrutierung.

Ein Mangel an Weiterbildungsmöglichkeiten auf dem Land hat ebenfalls eine negative Auswirkung auf die Rekrutierung und Erhaltung von Landärzten (Dussault & Franceschini, 2006). Studien (z.B. Kolstad, 2011) betonen die Wichtigkeit von Weiterbildungs- und Karrieremöglichkeiten für Ärzte bei der Entscheidung, wo sie arbeiten wollen, insbesondere bei der Rekrutierung. In einer Umfrage unter Krankenpflegern im australischen Bundesstaat Queensland wird dies als einer der wichtigsten Gründe für das Verlassen von ländlichen Regionen genannt, wodurch auch die Relevanz von Weiterbildungs- und Karrieremöglichkeiten für die Erhaltung von Landärzten angedeutet wird (Hegney et al., 2002). Ferner können sich fehlende Weiterbildungs- und Karrieremöglichkeiten auch negativ auf die Qualität der Versorgung von ländlichen Patienten auswirken (Herd et al., 2017; Rowe & Hargrave, 2014).

Systemische und institutionelle Faktoren

Als systemische Gründe werden hier alle jene Gründe gezählt, die aufgrund der Struktur des Ausbildungs-, Gesundheits- oder anderer Systeme und ihrer Institutionen, Ärzte von einer Niederlassung in ländlichen Gebieten abhalten.

Die Ausbildung in ländlichen Gebieten hat, ähnlich wie die individuellen Faktoren der „ländlichen Herkunft“, eine starke Auswirkung auf die Wahrscheinlichkeit eines Arztes als Landarzt tätig zu werden. Wie bereits vorhergehend beschrieben, spielt dort der Ort der Ausbildung eine Rolle für die individuellen Faktoren. Allerdings gibt es auch hier das systemische Pendant: So bildet ein kleiner Teil der medizinischen Fakultäten in den USA überproportional viele Landärzte aus (Rosenblatt, Whitcomb, Cullen, Lishner, & Hart, 1992). Kongruent mit dem Entwicklungsmodell von Hancock et al. (2009) handelt es sich hierbei häufig um Fakultäten im ländlichen Raum. Des Weiteren zeigt auch die Lehre in vielen Ländern einen starken Fokus auf die Arbeit im Krankenhaus und solche Spezialisierungen, welche ein Krankenhausumfeld benötigen. Das gilt sowohl weltweit (Dussault & Franceschini, 2006) als auch speziell für Deutschland (Holst et al., 2015). Außerdem decken häufig die Fortbildungssysteme Landärzte zu einem geringeren Grad ab als ihre Kollegen aus der Stadt (Dussault & Franceschini, 2006). In deutschen Studien spielt dieser Punkt auch eine Rolle, wenn auch eine weniger Wichtige (Kasch et al., 2016; Roick et al., 2012).

Zusätzlich ist das Krankenversicherungssystem in der Verfügbarkeit von Landärzten wichtig, da es über die Festsetzung und Regulierung der Vergütung die finanzielle Attraktivität maßgeblich gestaltet. Weltweit finden sich in einigen Ländern komplett getrennte Versorgungsysteme; so existiert neben einem staatlichen Versorgungs- und Krankenversicherungssystem auch ein rein privates System auf Selbstzahlerbasis mit einer deutlichen geringeren Regulierung, wie z.B. in Thailand (el el Koussa, Atun, Bowser, & Kruk, 2016). Daran lässt sich gut aufzeigen, dass Ärzte im privaten System ein deutlich höheres Einkommen erzielen können und deswegen häufig nicht mehr im staatlichen System praktizieren (Wibulpolprasert & Pengpaibon, 2003). Am

Beispiel von Thailand wird dadurch die Ungleichheit der Versorgunglage beschrieben, da somit ärmere Bevölkerungsschichten sich eine private medizinische Behandlung nicht leisten können und nicht genug Kapazitäten im staatlichen System verfügbar sind. Die Lage in Deutschland ist, im Vergleich zu Thailand, nicht derartig angespannt, aber nichtsdestotrotz trägt das deutsche Krankenversicherungssystem auch teilweise zu einem Landärztemangel bei. In ihrer Studie untersuchen Sundmacher and Ozegowski (2016) die Ungleichverteilung von Ärzten in Deutschland nach Region, wobei ländliche oder ökonomisch schwache Regionen hier eine deutlich geringere Dichte an Ärzten haben. Das Zusammenspiel von Gesetzlichen Krankenversicherungen (GKV) und Privaten Krankenversicherungen (PKV) mit jeweils eigenen Vergütungsmechanismen schafft verschiedene Anreize für Ärzte. Die Behandlung von GKV-Versicherten kann nur mit einer sozialrechtlichen Zulassung, auch „Kassenzulassung" genannt, abgerechnet werden. Mit dem Erhalt einer Kassenzulassung verpflichtet sich der entsprechende Arzt, eine gewisse Stundenanzahl für GKV-Versicherte freizuhalten. In der restlichen Zeit steht es dem Arzt jedoch offen, ob er auch GKV-Versicherte oder nur PKV-Versicherte behandelt. Die Abrechnung der Leistungen mit der Kassenärztlichen Vereinigung (KV) wird anhand des jährlich verhandelten Einheitlichen Bemessungsmaßstabes (EBM) durchgeführt. Das Modell funktioniert so, dass nach einem ersten Besuch eines GKV-Versicherten für weitere Behandlungen innerhalb eines Quartals deutlich weniger Vergütung erhalten wird. Dem gegenüber steht die Abrechnung von PKV-Versicherten nach Gebührenordnung für Ärzte (GOÄ), welche alle Leistungen einzeln abrechnen lässt (Walendzik, Manouguian, Greß, & Wasem, 2009). Somit kann ein Arzt seine Vergütung maximieren, indem mehr PKV-Versicherte behandelt werden. Weil in der PKV nur Patienten mit hohem Einkommen, Selbstständige oder Beamte versichert werden können (Porter & Guth, 2012), ist die Konzentration von PKV-Versicherten in städtischen Gebieten deutlich höher als auf dem Land. Aufgrund einer historisch bedingten Abwesenheit des Beamtenstatus in den Gebieten der ehemaligen Deutschen Demokratischen Republik, verbunden mit einem geringeren Durchschnittseinkommen (Sundmacher & Ozegowski, 2016), ist die Zahl der PKV-Versicherten in den neuen Bundesländern besonders niedrig. Daher erklären Sundmacher and Ozegowski (2016) die Disparität in der Verteilung von Ärzten zwischen ländlichen und städtischen Gebieten z.T. durch die Verteilung von PKV-Versicherten. Kistemann et al. (2007) sehen aber auch einen Teil des Problems bei den KVen. Diese regulieren lediglich die Anzahl der Ärzte mit Kassenzulassung innerhalb eines Planungsbereichs. Wenn dieses Gebiet jedoch zu heterogen ist, finden sich häufig alle Ärzte mit Kassenzulassung im städtischsten Teil, so dass für die Bewohner der ländlicheren Teile wiederum eine schlechtere Versorgungslage besteht.

Ein zusätzlicher, in der obenstehenden Beschreibung implizit genannter Grund, ist ein Übermaß an Bürokratie (van den Bussche et al., 2010). Eine Niederlassung auf dem Land als Allgemeinmediziner ergibt aufgrund der geringeren Quote von PKV-Versicherten nur mit einer Kassenzulassung finanziell Sinn. Jedoch verursacht diese einen erhöhten bürokratischen Aufwand, welcher vom eigentlichen, kurativen Tätigkeitsfeld des Arztes ablenkt. Dadurch wird die Tätigkeit als Landarzt weniger attraktiv, was zunächst in der Rekrutierung negativ auffällt (van den Bussche et al., 2010).

Verwandt damit ist die Tatsache, dass ein höherer Grad der Dezentralisierung häufig zu einer größeren Ungleichverteilung beiträgt. So kann es, wie im Fall der USA, zu Konkurrenzsituationen zwischen den Bundesstaaten kommen, in denen einzelne Bundesstaaten anderen durch Förderprogramme Ärzte abwerben (de Vise, 1973). Besonders ländlich geprägte Staaten haben häufig nicht die Budgets, um mit anderen Bundesstaaten mitzuhalten und werden damit zu „Exporteuren" von Ärzten. Generell geht eine Dezentralisierung häufig mit geringeren Budgets und höherem administrativen Aufwand einher, was effektiven Maßnahmen oft hinderlich gegenüber steht (Dussault & Franceschini, 2006). Dies wird auch in australischen Studien sichtbar, die sich in der Regel auf Ebene der Staaten und Territorien beschränken und verschiedene Problemstellungen und Herangehensweisen aufzeigen (Han & Humphreys, 2006; Hegney et al., 2002).

Soziokulturelle Faktoren

Diese letzte Kategorie umfasst die deutlich breitere Gesamtheit von Faktoren aus den Gebieten der Wirtschaft, Politik, Gesellschaft und Geschichte, die auf den Landarzt einwirken (Dussault & Franceschini, 2006).

Einer der wichtigsten Faktoren für die Erhaltung von Landärzten ist die erfolgreiche Integration in die ländliche Gemeinschaft. Laut Hancock et al. (2009), dessen Analyse auf vorhergehende Studien aufbaut (Cutchin, 1997a, 1997b), handelt es sich um die Integration sowohl in die Personengruppe an sich, z.B. in der Form von Freundschaften, als auch in die mit ihr verbundene Kultur, z.B. in der Form von Festen. Letzteres wird damit speziell als soziokulturelle Integration betitelt. Der Fokus gilt dort dem Engagement des Landarztes (Hancock et al., 2009). Hierbei gilt es einen Bezug zu dem Konzept der Ländlichkeit aufzubauen. Diese Idee ist eng verwandt mit dem Pfad „Sinn für den Ort" (orig. „sense of place") im zuvor beschriebenen Entwicklungsmodell von Hancock et al. (2009). Der Sinn für den Ort rührt vom Wunsch her, an einem Ort zu leben, dem man sich verbunden fühlt und durch den man Inspiration erfährt. Bei dem Begriff „Ort" gilt es jedoch vom rein geografischen Verständnis zu abstrahieren und auch soziokulturelle und historische Einflüsse miteinzubeziehen, welche auf den Landarzt einwirken (Cutchin, 1997b; Hancock et al., 2009). Über die Zeit verstärkt sich meistens dieser Effekt und sorgt für eine hohe Erhaltungsquote, da der Ort für den Landarzt identitätsstiftend wirkt und ihm das Gefühl von Sicherheit und Freiheit vermittelt. Diese Gefühle können auf persönlicher (z.B. Absicherung der Existenz der Familie durch Arbeit und ländliche Gemeinschaft) und professioneller Ebene (z.B. Ausbau und Verbesserung der medizinischen Versorgung) auftreten (Cutchin, 1997b). Insbesondere der Aspekt der Ve-

rantwortung der medizinischen Versorgung für die Gemeinschaft ist auch in Deutschland ausschlaggebend (Pohontsch, Hansen, Schäfer, & Scherer, 2018).

Im erweiterten Sinne lässt sich dieses Kriterium der Integration auch auf den Partner oder die Partnerin anwenden (Han & Humphreys, 2006). Mayo and Mathews (2006) zeigen ähnliche Ergebnisse und heben den Stellenwert einer Integration des Partners hervor. Die häufigsten Formen der Integration des Partners sind neben dem Arbeitsplatz die persönlichen Beziehungen zu Freunden. Besonders in dem Fall, dass Familie bereits in der ländlichen Region lebt, kommt ein hoher Stellenwert der Integration des Partners zu. Dies unterstreicht an dieser Stelle noch einmal die höhere Wahrscheinlichkeit von Ärzten mit ländlicher Herkunft als Landarzt zu praktizieren. Dem gegenüber steht das Fehlen der Integration, nämlich die Entfremdung, welche in manchen Studien als ein Grund, die ländliche Praxis zu verlassen, genannt wird (Hegney et al., 2002).

Das Prestige auf dem Land zu praktizieren, wird häufig als geringer wahrgenommen und erschwert dadurch die Rekrutierung von Landärzten (van den Bussche et al., 2010). Dies gilt besonders im Vergleich zu Kollegen, die in einem spezialisierten, städtischeren oder Krankenhauskontext arbeiten (Dussault & Franceschini, 2006). Prestige wird als wichtiger und relevanter Faktor in der beruflichen Laufbahn, besonders bei Ärzten ohne Partner und Kinder, beschrieben (Kasch et al., 2016).

Obendrein können die Freizeitangebote auf dem Land als weniger attraktiv wahrgenommen werden (Dussault & Franceschini, 2006). So sind manche sozialen und kulturellen Einrichtungen (z.B. Museen, Theater) in ländlichen Regionen nicht vorhanden und können die Rekrutierung, aber auch die Erhaltung von Landärzten, denen solche Angebote wichtig sind, erschweren. Dionne, Langlois, and Lemire (1987) untersuchen dies in ihrer Studie, bei welcher sie die Zahl und Bewertung der Restaurants stellvertretend für die Freizeitangebote verwenden. Diese Variable konnte eine im Vergleich zum Gehalt geringe, aber trotzdem nicht unerhebliche Varianz in der Ungleichverteilung von Ärzten in Quebec erklären.

Die Entscheidung zu bleiben oder zu gehen

All diese Faktoren beeinflussen die Rekrutierung und Erhaltung von Ärzten. In der Realität treten diese Faktoren selten isoliert auf, sondern verbinden sich zu komplexeren Gründen, in der ländlichen Praxis zu bleiben oder diese zu verlassen. Hancock et al. (2009) bilden diese in ihrer Studie ebenfalls ab. In einer Einzelfallentscheidung werden dann Gründe zu bleiben, wie z.B. Komfort, Gemeinschaft oder das Gefühl, etwas Positives zu tun, gegen Gründe zu gehen, etwa finanzielle Probleme oder der Wunsch nach Veränderung, abgewogen. Dies ist ein Prozess, welcher über längere Zeit passiert und die Erhaltung von Landärzten betrifft. Zusätzlich existieren gewisse Auslöser, die eine Entscheidung erzwingen. Zu diesen gehören negative Ereignisse, wie der Ruin der eigenen Praxis oder Verlust des Partners (Hancock et al., 2009).

2.2. Auswirkung des Landärztemangels

Die vorherbeschriebenen Faktoren beeinflussen die Rekrutierung und Erhaltung entweder auf einer individuellen oder auf einer systemischen Ebene direkt. Das folgende Unterkapitel schließt daran an und zeigt die Folgen eines Landärztemangels auf der Ebene des Gesundheitssystems auf. Der Fokus liegt hierbei insbesondere auf der medizinischen Grundversorgung, die viele Landärzte in ihrer Funktion als Hausärzte leisten. Eine Vielzahl an Studien untersucht statt einem Mangel an Landärzten einen Mangel an Hausärzten. Aufgrund der in Unterkapitel 1.3 gegebenen Definition eines Landarztes als Hausarzt im ländlichen Gebiet, lassen sich diese Erkenntnisse daher zum Großteil übertragen.

Landärzte stellen einen wichtigen Teil der medizinischen Grundversorgung dar (Starfield et al., 2005), so dass ein Mangel die gesamte medizinische Versorgung direkt negativ beeinflusst. Aufgrund des verringerten Angebots an medizinischen Versorgungsleistungen durch den Mangel an Landärzten wird die Quantität der verfügbaren medizinischen Versorgungsleistungen im Gesundheitsmarkt verringert.

Hierbei ist die Grenze zwischen einer akzeptablen Versorgung und einer Unterversorgung oft fließend (Dubben, Muche-Borowski, & Scherer, 2016). Dieser Umstand liegt in der Natur des Konsums und der Bereitstellung von medizinischen Versorgungsleistungen, welche jedoch stark abhängig vom jeweiligen Gesundheits- und Krankenversicherungssystem ist. So sind häufig die Kosten der medizinischen Versorgungsleistungen für den Patienten geringer als die eigentlichen Kosten, da diese über die Krankenversicherung subventioniert werden. Dieser Umstand kann zu einem Überkonsum führen – ein Verhalten, welches in der Literatur als „ex-post moral hazard“ bekannt ist (Arrow, 1963; Cutler & Zeckhauser, 2000; Zeckhauser, 1970). Der „ex-post moral hazard“ von Patientenseite betrifft insbesondere Landärzte, da sie als häufig einziger Arzt in ihrer Region der alleinige Ansprechpartner für ihre Patienten sind (Starfield et al., 2005). Zusätzlich ist die Landbevölkerung im Durchschnitt häufiger GKV versichert (Sundmacher & Ozegowski, 2016), was zu einer stärkeren Inanspruchnahme von Hausärzten im Vergleich zu Spezialisten führt und den Druck auf die verminderte Quantität der medizinischen Versorgungsleistungen zusätzlich erhöht (Bremer, P. & Wübker, A., 2013).

Die Qualität der medizinischen Versorgungsleistungen steht in einer komplexen Wechselbeziehung zu der Quantität dieser Leistungen und wird daher auch vermindert. Als Qualität wird in diesem Sinne der Behandlungserfolg, bzw. dessen Wahrscheinlichkeit, die Angemessenheit und Vorgehensweise des Diagnose- und Behandlungsprozesses sowie die Lebensqualität des Patienten mit allen ihren Aspekten verstanden (Donabedian, 1966). Die Verringerung der Qualität durch eine reduzierte Quantität hat eine Vielzahl an Gründen. So werden Patientenbesuche kürzer und die Zeit pro Patient sinkt, da somit die Anzahl an Patienten gesteigert und der Einbruch in der Quantität zumindest z.T. aufgefangen werden kann–wodurch jedoch die Qualität verringert

wird (Chen, Farwell, & Jha, 2009). Auch sinkt die Frequenz der Patientenkontakte, da Patienten länger auf einen Arzttermin oder andere medizinische Leistungen warten müssen (Bodenheimer & Pham, 2010). Ferner steigt durch die geringere Quantität die Belastung der verbliebenen Ärzte (Cooper, Rout, & Faragher, 1989). Dadurch sinkt die Arbeitszufriedenheit dieser Ärzte, was einen weiteren negativen Einfluss auf die Qualität der medizinischen Versorgungsleistungen haben kann (Grembowski et al., 2005; Mokkink et al., 1985). Eine geringere Arbeitszufriedenheit vermindert zusätzlich die Attraktivität des Berufs eines Landarztes, wodurch die Rekrutierung und Erhaltung wie in Unterkapitel 2.1 beschrieben verringert werden.

Eine reduzierte Qualität der medizinischen Versorgung führt zu einer höheren Sterblichkeit. Traditionell werden in der Literatur verschiedene Sterblichkeitsquoten gemessen und können so miteinander verglichen werden. In der Betrachtung für die USA fällt so eine höhere Sterblichkeit und niedrige Lebenserwartung in Regionen mit weniger Hausärzten, auch unabhängig von einem Einkommenseffekt, auf (Shi, Starfield, Kennedy, & Kawachi, 1999). Dies wurde auch für Deutschland nachgewiesen (Rau & Schmertmann, 2020). Da es sich hierbei zum Großteil um ländliche Gebiete handelt, lässt sich hier gut der Hausärztemangel im „Spezialfall“ als Landärztemangel zeigen. Darüber hinaus wird eine geringere Anzahl an Landärzten und damit eine geringere Dichte an medizinischer Grundversorgung mit einer höheren Kindersterblichkeit sowie einem geringeren Gewicht bei der Geburt assoziiert (Shi et al., 2004), einer für die Gesundheit des Kindes wichtigen Größe (McCormick, 1985).

Zusätzlich haben Quantität und Qualität der medizinischen Grundversorgung einen nichtunerheblichen Effekt auf Auftreten und Behandlungserfolg von sogenannten Volkskrankheiten. Macinko, Starfield, and Shi (2007) fassen eine Vielzahl an Studien, die die USA und deren medizinische Grundversorgung durch Ärzte untersuchen, zusammen. Dabei sticht heraus, dass eine geringere Dichte an Hausärzten mit einer höheren Anzahl an Schlaganfällen sowie einer späteren Erkennung von bestimmten Krebsarten assoziiert wird. Ähnliche Korrelationen wurden auch für Diabetes mellitus, Lungenentzündungen, Asthma und weitere Volkskrankheiten nachgewiesen (Guanais & Macinko, 2009; Parchman & Culler, 1994). Die Liste von weiteren Gesundheitsergebnissen, die durch einen Haus- bzw. Landärztemangel hervorgerufen werden können, ist lang. Weitere Studien beschreiben schlechtere Zahnhygiene, verringerte Sehkraft, unvollständige Impfungen oder gefährlichen Bluthochdruck als mögliche Folgen (Shi & Starfield, 2000). Hierbei ist insbesondere die Funktion des Hausarztes als erste Anlaufstelle und langfristiger Beziehungspunkt für medizinische Angelegenheit wichtig. So bietet der Hausarzt eine Möglichkeit, lebensstilbezogene Frage zu stellen und ungesundes Verhalten zu korrigieren (Whitlock, Orleans, Pender, & Allan, 2002). Dadurch haben Patienten frühzeitig die Chance, ungesundes Verhalten zu korrigieren und Folgeerkrankungen vorzubeugen. Darüber hinaus gilt, dass je häufiger der Haus- bzw. Landarzt aufgesucht wird, desto besser befolgen die Patienten medizinische Ratschläge (Heszen-Klemens & Lapińska, 1984).

Ein weiterer Effekt, der auch von Parchman and Culler (1994) beschrieben und auch an anderen Stellen nachgewiesen wird (z.B. Guanais & Macinko, 2009; Macinko et al., 2010), ist der einer höheren Rate an vermeidbaren Krankenhausaufenthalten bei einer geringeren Dichte an Haus- bzw. Landärzten. Dies verursacht zusätzliche Kosten für das Gesundheitssystem und blockiert weitere Ressourcen für nicht vermeidbare Krankenhausaufenthalte. In dem Fall Brasiliens werden Einsparungen von bis zu 63 Millionen US Dollar[2] für 87,7 Millionen Brasilianer über den Verlauf von 3 Jahren errechnet (Guanais & Macinko, 2009). Dies entspricht jährlichen Einsparungen von knapp 240000 US Dollar pro Kopf. Auch wenn nicht der komplette Teil der Einsparungen durch eine Verbesserung der Grundversorgung erklärt werden konnte, liegt das Bestimmtheitsmaß R^2 abhängig vom Krankheitsbild in einem Bereich von 73% bis 87% und weist somit eine hohe Güte auf.

Von einem Landärztemangel werden sozial schwache Gruppen auf dem Land besonders getroffen. Allein schon durch einen Mangel an Hausärzten steigt die Ungleichheit an Gesundheitsergebnissen innerhalb der Bevölkerung, da somit die Versorgung speziell für all jene verschlechtert wird, die besonders von Hausärzten abhängig sind. Dazu gehören Ältere und chronisch Kranke (Bodenheimer & Pham, 2010). Diese benötigen vor allem eine höhere Frequenz als auch mehr Zeit pro Besuch (Blumenthal, Culpepper, Stafford, & Starfield, 1999), so dass das verringerte Angebot an medizinischen Versorgungsleistungen hier einen größeren negativen Effekt hat. Ferner benötigen auch weniger Gebildete, Arbeitslose und Sozialhilfeempfänger Hausärzte zu einem größeren Teil als der Durchschnittsbürger (Luppa, Giersdorf, Riedel-Heller, Prütz, & Rommel, 2020). Infolgedessen leiden auch diese besonders stark unter einem Mangel. Zwischen Land und Stadt findet sich vor allem ein Gefälle auf Basis des Alters und des Verdienstes, wobei auf dem Land häufiger Ältere und weniger Gutverdienende gefunden werden (Sundmacher & Ozegowski, 2016), die durch den Landärztemangel besonders benachteiligt werden.

2.3. Anreizsysteme im internationalen Vergleich

In diesem Unterkapitel wird eine Übersicht über bestehende Anreizsysteme in anderen Ländern gegeben. Da neben Deutschland auch andere Länder von einem Landärztemangel betroffen sind, findet sich im internationalen Vergleich eine Vielzahl von verschiedenen Anreizsystemen. Häufig sind es regulatorische und administrative Maßnahmen, finanzielle und professionelle Anreize oder eine Mischung von beiden. Fruen et al. (1980) sehen, im Gegensatz zu größeren strukturellen Anpassungen, einen Vorteil solcher Programme darin, dass eine Ungleichheit zielgenau ausgeglichen werden kann. Es müssen nicht Anreize für alle Ärzte geschaffen

[2] Circa 59,9 Millionen Euro zum Zeitpunkt der Veröffentlichung der Studie (Europäische Zentralbank, 2020)

werden, sondern nur für eine ausreichende Anzahl, um den Landärztemangel zu beheben.

Regulatorische und administrative Maßnahmen

Unter die Kategorie der regulatorischen und administrativen Maßnahmen fallen Verträge, Verordnungen oder andere staatliche Maßnahmen, welche die Arbeit in ländlichen Gebieten meistens für einen bestimmten Zeitraum vorschreiben (Dussault & Franceschini, 2006).

Australien gibt Ärzten, die nicht ebenda ihre Ausbildung abgeschlossen haben, die Möglichkeit, ihre australische Approbation nach einer verpflichtenden Landarzttätigkeit von fünf Jahren zu erhalten (Han & Humphreys, 2006). Alternativ müssten sie für die australische Approbation ein weiteres Studienjahr und eine Abschlussprüfung absolvieren. Dadurch wird ein Teil der ländlichen Versorgung sichergestellt, allerdings mit einigen Einschränkungen: So verlässt ein Großteil dieser internationalen Ärzte die ländliche Praxis nach Ablauf der fünf Jahre, so dass die ländliche Versorgung auf einen immer neuen Zustrom von neuen internationalen Ärzten angewiesen ist.

Andere Länder (z.B. Ghana und Thailand) haben ihre medizinische Ausbildung insofern modifiziert, als dass eine verpflichtende Station die Arbeit auf dem Land ist (Dussault & Franceschini, 2006). Wie bereits vorhergehend beschrieben, kann die Ausbildung auf dem Land dazu beitragen, dass sich die angehenden Ärzte eher in einem ländlichen Gebiet niederlassen. In Indonesien, Vietnam und Südafrika muss im Rahmen eines Medizinstudiums eine gewisse Zeit verpflichtend im öffentlichen Dienst gearbeitet werden. Dies nutzen die Länder, um damit die Versorgung in ländlichen Gebieten zu verbessern (Dussault & Franceschini, 2006). Auch in Lateinamerika gibt es einige Länder mit ähnlichen Vorgehensweisen, wie z.B. das *Mais Médicos*-Programm in Brasilien (Pereira, Santos, Santos, Oliveira, & Rattner, 2016). Hier sind jedoch die verpflichteten Ärzte durch eine besonders geringe Produktivität aufgrund fehlender Motivation gekennzeichnet (Dussault & Franceschini, 2006).

In Thailand gibt es weitergehende Maßnahmen, die nochmals stärker in den Ausbildungsprozess eingreifen. So wurden die Prozesse der Zulassung zum Medizinstudium reglementiert und die Auswahl einer Spezialisierung von der thailändischen Regierung limitiert. Wer sich verpflichtet, auf dem Land zu praktizieren, wird garantiert zum Studium zugelassen und kann sich eine Spezialisierung frei aussuchen (Wibulpolprasert & Pengpaibon, 2003). Auch wenn diese Maßnahme zu einer Verbesserung der Versorgung auf dem Land beigetragen hat, wird sie von der medizinischen Fachwelt aufgrund der Tiefe des Eingriffs in die Selbstbestimmung von Ärzten scharf kritisiert (Dussault & Franceschini, 2006). Darüber hinaus wurden Programme aufgesetzt, welche speziell Bewohner ländlicher Gebiete für ein Medizinstudium rekrutieren, da diese eine höhere Wahrscheinlichkeit haben, in ländlichen Gebieten zu praktizieren (Wibulpolprasert & Pengpaibon, 2003). Auch in der kanadischen Provinz Quebec wurden ähnliche Maßnahmen implementiert (Dussault & Franceschini, 2006).

Vielen dieser Programme ist gemein, dass sie häufig nur begrenzt erfolgreich sind (Dussault & Franceschini, 2006). Zwar resultieren sie in einer höheren Zahl von Ärzten auf dem Land, allerdings häufig zum Preis von geringerer Produktivität und/oder Motivation. Auch wird ersichtlich, dass ihr Nutzen begrenzt ist, wenn die restliche medizinische Infrastruktur (z.B. medizinische Geräte) nicht vorhanden ist. Ferner werden sie oft als „unfair" charakterisiert (Dussault & Franceschini, 2006) und stellen durch ihren Zwangscharakter im internationalen Kontext einen weitgehenden Eingriff in die Freiheit der medizinischen Fachwelt dar, was zu einer Emigration von Ärzten führen kann (Dussault & Franceschini, 2006).

Finanzielle und professionelle Anreize

Eine andere Möglichkeit ist es, Ärzten positive Anreize zu bieten, um auf dem Land zu praktizieren. Neben verschiedenen finanziellen Möglichkeiten können auch professionelle Anreize für den Arzt oder Partner gesetzt werden. Studien beschreiben bei Letzteren eine breite Auswahl wie z.B. Mentoringprogramme, flexiblere Arbeitszeiten oder mehr Freiheit bei der Urlaubsgestaltung (Dussault & Franceschini, 2006). Für die finanziellen Anreize bestehen auch viele Möglichkeiten. Die genaue Struktur dieser ist von der jeweiligen Jurisdiktion (i.d.R. Staat; in den USA, Kanada oder Australien auch Bundesstaaten) und der dortigen Struktur des Ausbildung- und Gesundheitssystems abhängig.

In den USA gibt es im Moment, und gab es in der Vergangenheit viele Programme, welche auf rein finanzielle Anreize abzielen. Diese finden z.T. auf Bundesebene („federal level") und z.T. auf Ebene der Bundesstaaten („state level"), statt. So wurden bedingte Stipendien vergeben, bei denen sich der Empfänger verpflichtet, für einen begrenzten Zeitraum auf dem Land zu praktizieren. Andere Formen sind Kredite mit einer ähnlichen Verpflichtung, Studienkreditrückzahlungsprogramme oder direkte Zahlungen an Landärzte (Daniels et al., 2007). Übliche Variablen dieser Programme sind hier Zahlungszeitpunkt und -dauer sowie weitere Modalitäten zu Ausfall, Kündigung und Stundung (Bärnighausen & Bloom, 2009). Teilweise werden weitere Leistungen wie Mentoringprogramme oder Praktika miteingebunden (Dussault & Franceschini, 2006). Manche dieser Programme konnten in den USA Erfolge vorweisen. Trotz zahlreicher Studien herrscht keine Einigkeit darüber, weshalb manche Programme deutlich erfolgreicher sind als andere.

Auch Australien arbeitet mit ähnlichen Programmen, die hauptsächlich auf finanzielle Anreize setzen und dabei deren Begrenzung aufzeigen. Hier finden sich viele Zuschüsse, wie z.B. für den Umzug aufs Land, die Ausstattung der Praxis oder die weitere Ausbildung (Scott et al., 2013). Das Problem der Erhaltung von Landärzten wird auch z.T. mit Zuschüssen, sogenannten „retention grants", versucht zu lösen (Han & Humphreys, 2006). Jedoch hat sich im australischen Kontext auch die Begrenzung rein finanzieller Anreize gezeigt. Scott et al. (2013) haben in einem Discrete-Choice-Experiment untersucht, wie hoch die finanziellen

Anreize ausfallen müssten. Abhängig vom Standort auf dem Land (Inland vs. Küste) und den anfallenden Aufgaben kamen sie auf einen einmaligen Anreizbetrag von bis zu 237000 australische Dollar[3], was in etwa 230% des durchschnittlichen Jahreseinkommens eines australischen Arztes entspricht (Scott et al., 2013). Ähnliche Größenordnungen treten auch bei anderen Studien auf.

Im Gegensatz zu bestehenden Ärzten, scheinen jedoch bei Ärzten kurz nach der Approbation finanzielle Anreize wirkungsvoller zu sein. So zeigte eine weitere australische Studie die Wirksamkeit finanzieller Anreize für die Rekrutierung junger Landärzte (Yong, Scott, Gravelle, Sivey, & McGrail, 2018). Insbesondere bei einer großen Auswahl an möglichen Orten für die Tätigkeit als Landarzt ist der Effekt stark.

In der kanadischen Provinz Quebec erhalten Landärzte höhere Erstattungssätze als ihre Kollegen in der Stadt, sowie weitere finanzielle Unterstützung beim Aufbau der Praxis (Dionne et al., 1987). Zwar stellen Erstattungssätze nur einen Hebel in der Gewinngleichung einer Praxis dar, allerdings kann hier eine kleine Anpassung schon weitreichende Auswirkungen für die Vergütung des Arztes haben (Fruen et al., 1980). Dieses Programm, das 1982 eingeführt wurde, ist im internationalen Vergleich als relativ erfolgreich bewertet worden (Bolduc, Laval, & Fortin, 1996). Zusätzlich wurden die Landärzte in Quebec jedoch mit garantierten Weiterbildungsmaßnahmen unterstützt, was in einer Umfrage als wichtig beschrieben wurde (Dussault & Franceschini, 2006).

Thailand ergänzt seine administrativen und regulatorischen Maßnahmen mit weiteren finanziellen Anreizen. So erhalten Ärzte pauschale Zahlungen für die Arbeit in ländlichen Gebieten. Auch werden zusätzliche Zahlungen dafür geboten, keine private Praxis zu eröffnen bzw. die Aktivitäten dieser zu beschränken. Hintergrund dafür ist, dass für thailändische Ärzte eine weitere, private Praxis deutlich profitabler ist, wo jedoch nur gut zahlende Patienten behandelt werden, was die Versorgunglage belastet (Dussault & Franceschini, 2006). Des Weiteren vergibt die thailändische Gesellschaft von Landärzten („The Rural Doctor Society“) Auszeichnungen und Ehrendoktortitel für Landärzte, die sich besonders verdient gemacht haben und zielt damit auf den Prestigefaktor ab (Dussault & Franceschini, 2006).

Auf den Philippinen wurde auch ein rein finanzielles Programm aufgesetzt, welches nach seinem Auslaufen deutlich die Nachteile finanzieller Anreize zeigt. Fast sämtliche Ärzte, die durch das Programm in ländliche Gebiete gewechselt hatten, schlossen danach ihre Praxis auf dem Land und zogen zurück in städtische Gebiete (Dussault & Franceschini, 2006). Ein ähnliches Programm in Indonesien zeigt vergleichbare Effekte. Hier erhalten Ärzte, welche sich verpflichten, auf dem Land zu praktizieren, deutlich höhere Löhne und nach dem Ablauf einer Dreijahresfrist versprochene Karrieremöglichkeiten. Auswertungen zeigen, dass der Großteil der Teilnehmer kurz nach dem Ablauf der drei Jahre ihre ländliche Praxis schließt und in die Stadt zurückkehrt. Generell zeigen die Programme aus Indonesien und den Philippinen das Problem dieser finanziellen Anreize: Sie funktionieren nur, solange ausreichend Geld fließt (Bärnighausen & Bloom, 2009).

Es lässt sich kein abschließendes Resumé ziehen, ob regulatorische und administrative Maßnahmen oder finanzielle und professionelle Anreize die bessere Möglichkeit darstellen, um einen Landärztemangel zu beheben.

3. Methodik

Für diese Arbeit wurde ein Discrete-Choice-Experiment (DCE) als methodische Vorgehensweise gewählt. Bei einem DCE werden Daten innerhalb einer Umfrage erhoben und können vereinfacht ausgewertet werden, da die relevanten Faktoren in einem kontrollierten Umfeld experimentell manipuliert werden können. Anstelle von Bewertungen im Rahmen von Rating- oder Rankingfragen zu sammeln, werden die Probanden gebeten, sich wiederholt zwischen verschiedenen Angeboten zu entscheiden. Jeder Proband sieht hierbei die gleichen Auswahlentscheidungsblöcke („choice sets“) mit den gleichen Entscheidungsalternativen („alternatives“). Diese Alternativen stellen Szenarien dar, die sich durch ihre Eigenschaften („attributes“) und dem Grad der Ausprägung dieser Eigenschaften („levels“) unterscheiden. Diese Vorgehensweise schafft es, Situationen zu simulieren, welche nah an den tagtäglichen Entscheidungen von Ärzten, Patienten und weiteren Akteuren des Gesundheitssystem liegen. Daher werden sie bevorzugt in der akademischen Forschung von Gesundheitssystemen angewandt, um Präferenzen oder Kompromissentscheidungen zu untersuchen (Ryan et al., 2008). Darüber hinaus bauen DCEs auf vorhandene Erkenntnisse der Soziologie und der Verhaltensforschung auf (Swait & Andrews, 2003). Tiefergehend liegt die Random Utility Theory zugrunde, so dass DCEs sich in die Tradition dieser langjährig bekannten und erprobten Theorie einordnen lassen. Eine Literaturübersicht bezüglich DCEs im Bereich der Gesundheitsökonomie kann bei Clark, Determann, Petrou, Moro, and de Bekker-Grob (2014) gefunden werden.

Um die Präferenzen von Medizinstudenten für die Arbeit als Landarzt zu untersuchen, wurde im DCE ein Programm entworfen, bei dem sich die Probanden gegen die Zahlung einer monatlichen Geldsumme während des Studiums verpflichten, im Anschluss daran in einer vorher definierten Stelle auf dem Land zu arbeiten. Somit handelt es sich bei dieser Arbeit um die Erste ihrer Art, da infolgedessen Präferenzen bezüglich Eigenschaften von Landarztstellen quantifiziert werden konnten. Damit werden bestehende DCEs zum Landärztemangel mit der Arbeit von Yong et al. (2018) bezüglich der höheren Wirksamkeit finanzieller Anreize bei jungen Ärzten verbunden.

3.1. Untersuchte Eigenschaften

Zur Bestimmung der Eigenschaften und ihrer zugehörigen Ausprägungen wurde im ersten Schritt ein tiefgehender,

[3] Circa 171000 EURO zum Zeitpunkt der Veröffentlichung der Studie (Europäische Zentralbank, 2020)

semistrukturierter Literaturüberblick durchgeführt. Daraus wurde eine weite Spannbreite an möglichen Eigenschaften identifiziert. Um jedoch die Komplexität des DCEs nicht zu stark zu erhöhen, gilt es die Anzahl der Eigenschaften zu begrenzen. Die spätere Analyse basiert auf der Annahme, dass die Probanden Abwägungen zwischen den einzelnen Eigenschaften treffen, mit dem Ziel, ihren Nutzwert zu maximieren. Zu viele Eigenschaften könnten durch eine hohe Komplexität der Auswahlentscheidungsblöcke dazu führen, dass die Probanden stattdessen einfache Heuristiken anwenden (Witt, Scott, & Osborne, 2009). Daher wurde die Zahl der Eigenschaften auf sechs begrenzt.

Zur Validierung der daraus folgenden Eigenschaften und ihrer Ausprägungen wurde daraufhin eine Vorstudie als Onlineumfrage mit der relevanten Zielgruppe, Medizinstudenten in Deutschland, durchgeführt. Im Zeitraum von sechs Tagen nahmen 31 Medizinstudenten teil, von denen der Großteil zwischen 18 und 25 Jahre alt war (84%) und sich in der Vorklinik oder Klinik befanden (ebenfalls 84%). Diese Umfrage startete mit den folgenden Eigenschaften: (I) Ort der ersten Anstellung nach dem Studium, (II) Nettovergütung während des Studiums, (III) Flexibilität der Arbeit, (IV) Arbeitsstätte, (V) Verpflichtende 24h Hausbesuche und Bereitschaftsdienste und (VI) Länge der Verpflichtung. Zuerst wurden die Probanden gebeten, diese Eigenschaften nach Wichtigkeit zu ordnen, wobei (I) Ort der ersten Anstellung nach dem Studium, (II) Nettovergütung während des Studiums und (VI) Länge der Verpflichtung als wichtigste Eigenschaften gewählt wurden. Die Eigenschaft (V) Verpflichtende 24h Hausbesuche und Bereitschaftsdienste wurde durchgängig als unwichtig eingeordnet. In der darauffolgenden Freitextfrage nach unwichtigen Eigenschaften unter Angabe einer Begründung, wurden die verpflichtenden 24h Hausbesuche und Bereitschaftsdienste in 18 von 31 Fällen genannt. Begründet wurde das von den Probanden dadurch, dass sie sich „jung“ und „fit“ genug für Hausbesuche und Bereitschaftsdienste fühlen und, dass diese „Teil des Jobs“ seien. Aufgrund dessen wurde die Eigenschaft (V) Verpflichtende 24h Hausbesuche und Bereitschaftsdienste entfernt.

Weiterhin wurden die Probanden nach Eigenschaften gefragt, welche nicht unter den ursprünglichen sechs Eigenschaften genannt wurden, aber trotzdem für ihre Entscheidung wichtig wären. Neben Nennungen bereits angeführter Eigenschaften bzw. deren Konkretisierungen, wurden hauptsächlich die Themenbereiche der Weiterbildung (6 Nennungen), der Karrieremöglichkeiten und der Reputation (4 Nennungen) sowie der Vergütung nach dem Studium (3 Nennungen) angesprochen. Für die Auswahl einer weiteren Eigenschaft wurden alle drei Themenbereiche berücksichtigt. Da die Vergütung nach dem Studium bundesweit für viele Stellen durch Tarifverträge vorgeschrieben ist oder sich an diesen Tarifverträgen orientiert, wurde hieraus keine weitere Eigenschaft entworfen. Bezüglich der Karrieremöglichkeiten und Reputation trat die Schwierigkeit auf, dass mögliche Eigenschaften in ihrer Benennung und ihrer Ausprägungen subjektiv wahrgenommen werden können und darunter die Qualität der Ergebnisse leiden könnte. Neben den meisten Nennungen wurde der Aspekt der Weiterbildung als eine weitere Eigenschaft ausgewählt, da es bereits von einigen KVen, wie zum Beispiel in Baden-Württemberg, Programme gibt, bei denen junge Ärzte in Facharztweiterbildung unterstützt werden (KV Baden-Württemberg, 2018). Somit wurde für (V) Unterstützung bei der Facharztweiterbildung gewählt. Im Rahmen dieser Frage stellte eine Probandin die hohe Wichtigkeit der Möglichkeit, Teilzeit zu arbeiten heraus, um besser für ihr Kind sorgen zu können. Daher wurde (III) Flexibilität der Arbeit trotz der geringeren durchschnittlichen Wichtigkeit im Vergleich zu den anderen Eigenschaften belassen. Somit konnten im Rahmen der Vorstudie sechs Eigenschaften validiert werden. Tabelle 1 zeigt eine Übersicht der Eigenschaften und ihrer Ausprägungen. Zusätzlich werden diese auch nachfolgend beschrieben und kurz erläutert.

Die erste Eigenschaft „**Ort der ersten Anstellung nach dem Studium**“ beschreibt den Ort, an welchem der Medizinstudent nach dem Studium seine Arbeitsstätte hat. Folgend der Definition des Bundesamtes für Bauwesen und Raumordnung werden hier und im weiteren Verlauf dieser Arbeit Kleinstadt, Landstadt und Dorf anhand ihrer Einwohnerzahlen definiert. Ferner wird auch die Distanz zum nächsten Ballungsgebiet berücksichtigt (Stichwort: „Speckgürtel“), indem jeweils die Erreichbarkeit hinzugefügt wurde. Dafür werden die Abstufungen „ca. 30 min“ und „ca. 60 min“ verwendet, welche im bundesdeutschen Durchschnitt realistische Distanzen für Orte dieser Größe darstellen (Marschner et al., 2018; Marschner, Wasem, & Heymann, 2019). Aus der Größe des Wohnorts und der Erreichbarkeit von Ballungsgebieten wurden vier Kombinationen abgeleitet, welche die vier Ausprägungen dieser Eigenschaft darstellen.

Die zweite Eigenschaft „**Monatliche Nettovergütung während des Studiums**“ stellt die Vergütung dar, die während des Studiums an den Medizinstudenten gezahlt wird. Hierbei sind die Werte der verschiedenen Ausprägungen als monatliche Nettozahlungen in Euro zu verstehen. Die erste Ausprägung von 300 € ist äquivalent zum sogenannten „Büchergeld“, welches viele Begabtenförderwerke einkommensunabhängig an ihre Stipendiaten zahlen. Die zweite Ausprägung liegt mit 900 € etwas über dem aktuellen Höchstsatz der Förderung durch das Bundesausbildungsförderungsgesetztes (BAföG), der auch in etwa dem studentischen Bedarf in Deutschland entspricht (Oberlandesgericht Düsseldorf, 2020). Zusätzlich wurden weitere Ausprägungen in der Höhe von 1200 € , 1500 € und 2000 € definiert, welche über dem Durchschnitt des studentischen Bedarfs liegen. Für den Preisparameter wurde eine Vergütung während des Studiums gewählt, anstelle des Gehalts während der Tätigkeit als Arzt, da Letzteres eine Reihe an Komplikationen aufweist. In Deutschland wird die Vergütung für Ärzte entweder direkt durch Tarifverträge festgelegt oder sie orientiert sich zumindest stark daran. Ferner lässt sich durch eine Verpflichtung während des Studiums die Anzahl der Landärzte besser steuern und planen, da durch die Verpflichtung eine feste Zahl an praktizierenden Landärzten zu einem bestimmten Zeitpunkt in der Zukunft vorhanden

sein wird. Somit lässt sich dieses Verpflichtungsprogramm in den Bedarfsplanungsprozess der KBV integrieren.

Die dritte Eigenschaft „**Flexibilität der Arbeit**“ bildet die Flexibilität der Arbeit nach dem Studium ab. Hierbei sind die beiden Ausprägungen einerseits die Vollzeitarbeit und andererseits die Möglichkeit Teilzeit zu arbeiten.

Die **„Arbeitsstätte**“ wird in der vierten Eigenschaft dargelegt. Hierbei werden die folgenden drei Ausprägungen untersucht. Erstens kann der Medizinstudent nach seiner Approbation in einem Krankenhaus arbeiten. Zweitens ist eine Anstellung in einem medizinischen Versorgungszentrum (MVZ) möglich. Drittens besteht die Möglichkeit einer Anstellung in einer Praxis.

Als fünfte Eigenschaft wird „**Unterstützung bei der Facharztweiterbildung**“ angeführt. Diese ist an ähnliche Programme von einigen Landesärztekammern angelehnt. Voraussetzung für einen Facharzt ist neben anderen Anforderungen das Durchlaufen verschiedener Stellen und das erfolgreiche Abschließen einer Prüfung (Schmidt, 2008). Daher werden als Ausprägungen des Verpflichtungsprogramms folgende Unterstützungsmöglichkeiten aufgeführt: Erstens, keine Unterstützung bei der Facharztweiterbildung. Zweitens, administrative Unterstützung, indem z.B. ein Rotationsprogramm für die benötigten Stellen angeboten wird. Drittens, administrative und inhaltliche Unterstützung, bei welcher zusätzlich zum beispielsweisen Rotationsprogramm noch Mentoring angeboten wird, um den Lerneffekt zu maximieren.

Die sechste und letzte Eigenschaft „**Länge der Verpflichtung**“ ist die Länge der Verpflichtung, die der Medizinstudent mit diesem Arrangement eingeht. Die drei Ausprägungen umfassen hier jeweils eine Verpflichtung von einem, zwei oder drei Jahren. Dies ist absichtlich deutlich kürzer als ähnliche Verpflichtungsprogramme, z.B. jene der Bundeswehr, gewählt.

3.2. Studienaufbau

Der Fragebogen, welcher mit Hilfe des Online-Umfragetools DISE erstellt wurde (Schlereth, Eckert, & Skiera, 2012), besteht aus vier Abschnitten: Im ersten Abschnitt wurden studien- und berufswahlbezogene Daten erhoben, darunter Studienabschnitt, Semester und Universität. Somit soll sichergestellt werden, dass lediglich Medizinstudenten an der Umfrage teilnehmen. Zusätzlich wurden Präferenzen bezüglich der medizinischen Fachrichtung sowie des ersten Arbeitsortes erfragt. Zur Überleitung zum nachfolgenden DCE wurde am Ende des ersten Abschnitts die Einstellung gegenüber Verpflichtungsprogrammen für Medizinstudenten auf einer fünfstufigen Likertskala abgefragt.

Im zweiten Abschnitt erfolgte der Einsatz des DCEs. Das Choice-Design basiert auf den vorherbeschriebenen Eigenschaften und deren Ausprägungen. Die Eigenschaften sind voneinander unabhängig und die Ausprägungslevel sind ausgewogen aufgrund von minimalen Überschneidungen, da alle Eigenschaften und Ausprägungen in gleicher Häufigkeit erscheinen. Dies wurde durch die Generierung der dazugehörenden Matrix mithilfe der Software JMP sichergestellt. Die Gestaltung des DCEs kann aufgrund einer D-Effizienz von 95,5% als effizient beschreiben werden. Im Rahmen dieses Choice-Designs wurden insgesamt 12 Choice-Sets mit jeweils zwei Angeboten für das Verpflichtungsprogramm sowie einer No-choice Option gezeigt. Ein beispielhafter Auswahlentscheidungsblock („choice set“) ist in Abbildung 2 dargestellt.

Der dritte Abschnitt knüpft inhaltlich an den vorherigen an, indem er dem Probanden die Möglichkeit bietet, sich qualitativ zum DCE zu äußern. Die erste Freitextfrage fragt nach weiteren Überlegungen, die für den Probanden wichtig sind, um an dem im DCE beschriebenen Verpflichtungsprogramm teilzunehmen. Damit wird diese Linie aus der Vorstudie fortgeführt und zielt darauf ab, für zukünftige Studien Impulse zu geben. Die zweite Freitextfrage fragt nach Faktoren, die für die Fortführung der Arbeit als Landarzt nach Ablauf der Verpflichtung relevant sind. Mithilfe dieser soll skizziert werden, wie der Einfluss des Programms auf die medizinische Versorgung auf dem Land maximiert werden kann. Überdies sollen beide Fragen helfen, ein genaueres Bild von den Präferenzen zukünftiger Ärzte zu zeichnen.

Im vierten und letzten Abschnitt werden die demografischen Daten der Probanden erfasst. Neben Geschlecht, Alter und Wohnort wurden auch andere, thematisch spezifizierte Daten abgefragt. So wurden für das Thema Landarzt noch die Frage, wie viele Elternteile Ärzte sind, sowie die Frage nach dem Ort des Aufwachsens gestellt (Hancock et al., 2009). Die Umfrage schließt mit einem Aufmerksamkeitstest, bei dem der Proband gebeten wird, aus vier Optionen die Dritte auszuwählen.

3.3. Daten

Über den Verlauf von vier Wochen nahmen 345 Probanden an der Umfrage teil. Damit liegt die Zahl der Probanden deutlich über zahlreichen Mindeststichprobenempfehlungen, welche in der Literatur bezüglich DCEs gefunden werden können (de Bekker-Grob, Donkers, Jonker, & Stolk, 2015). Basierend auf der Faustregel von Johnson and Orme (2003) sowie Orme (1996) liegt die Mindeststichprobe bei 70 Probanden, die deutlich überschritten wird. Generell folgt diese Arbeit jedoch dem Ziel, eine möglichst große Stichprobe zu nutzen (de Bekker-Grob et al., 2015). Für die Auswertung mussten die Ergebnisse von zehn Probanden eliminiert werden. Davon wurden zwei durch den Aufmerksamkeitstest ausgeschlossen und acht weitere, da sie als Universität keine deutsche Medizinfakultät angegeben haben.

Die Probanden wurden über verschieden Kanäle akquiriert. Es wurden die Fachschaften und Studiendekanate aller 38 medizinischen Fakultäten öffentlicher Universitäten sowie der privaten Universität Witten/Herdecke kontaktiert. Zusätzlich wurden Probanden über persönliche sowie soziale Netzwerke (insb. Facebook) angeworben. Um Anreize für die Probanden zu schaffen, wurden fünf Gutscheine für einen Onlineshop im Wert von jeweils 25 € verlost.

Die Stichprobe ist nicht repräsentativ, bildet jedoch einen heterogenen Ausschnitt aus der Gesamtbevölkerung ab. Die 335 Probanden setzen sich aus 66,3% weiblichen,

Table 1: Eigenschaften des Discrete-Choice-Experiments und die zugehörigen Ausprägungen

Eigenschaft	Ausprägung 1	Ausprägung 2	Ausprägung 3	Ausprägung 4	Ausprägung 5	Ausprägung 6
(I) Ort der ersten Anstellung nach dem Studium	Landstadt (ca. 5000 Einwohner), ca. 30 min zum nächsten Ballungsgebiet	Landstadt (ca. 10000 Einwohner), ca. 60 min zum nächsten Ballungsgebiet	Dorf (ca. 1000 Einwohner), ca. 30 min zum nächsten Ballungsgebiet	Dorf (ca. 500 Einwohner), ca. 60 min zum nächsten Ballungsgebiet		Ich nehme keines der Angebote an, damit ich mir den Arbeitsort nach meinem Studium aussuchen kann.
(II) Monatliche Nettovergütung während des Studiums	300 €	900 €	1200 €	1500 €	2000 €	Ich nehme keines der Angebote an, damit ich mir den Arbeitsort nach meinem Studium aussuchen kann.
(III) Flexibilität der Arbeit	Vollzeit	Teilzeit möglich				Ich nehme keines der Angebote an, damit ich mir den Arbeitsort nach meinem Studium aussuchen kann.
(IV) Arbeitsstätte	Angestellt in einem Krankenhaus	Angestellt in einem MVZ	Angestellt in einer Praxis			Ich nehme keines der Angebote an, damit ich mir den Arbeitsort nach meinem Studium aussuchen kann.
(V) Unterstützung bei der Facharztweiterbildung	Keine Unterstützung bei der Facharztweiterbildung	Administrative Unterstützung bei der Facharztweiterbildung (z.B. Rotationsprogramm)	Administrative und inhaltliche Unterstützung bei der Facharztweiterbildung (z.B. Rotationsprogramm und Mentoring)			Ich nehme keines der Angebote an, damit ich mir den Arbeitsort nach meinem Studium aussuchen kann.
(VI) Länge der Verpflichtung	1 Jahr	2 Jahre	3 Jahre			Ich nehme keines der Angebote an, damit ich mir den Arbeitsort nach meinem Studium aussuchen kann.

Für welches Programm würden Sie sich am ehesten entscheiden?			
Ort der ersten Anstellung nach dem Studium	Landstadt (ca. 10.000 Einwohner), ca. 60 min bis zum nächsten Ballungsgebiet	Dorf (ca. 1.000 Einwohner), ca. 30 min bis zum nächsten Ballungsgebiet	
Monatliche Nettovergütung während des Studiums	300 Euro	1500 Euro	
Flexibilität der Arbeit	Teilzeit möglich	Vollzeit	Ich nehme keines der Angebote an, damit ich mir den Arbeitsort nach dem Studium selbst aussuchen kann.
Arbeitsstätte	Angestellt in einem MVZ	Angestellt in einer Praxis	
Unterstützung bei der Facharztweiterbildung	Administrative und inhaltliche Unterstützung (z.B. Rotationsprogramm und Mentoring)	Keine Unterstützung	
Länge der Verpflichtung	1 Jahr	2 Jahre	
	◯	◯	◯

Figure 2: Beispielhafte Darstellung eines Auswahlentscheidungsblocks

Quelle: Eigene Darstellung

33,4% männlichen und 0,3% diversen Medizinstudenten in Deutschland zusammen. Von diesen befinden sich 34,6% im ersten Studienabschnitt (Vorklinik), 51,6% im zweiten Studienabschnitt (Klinik) und 9,9% im Praktischen Jahr (PJ). 3,9% lassen sich hierbei nicht zuordnen, da sie entweder ein Urlaubssemester nehmen oder in einem Modellstudiengang eingeschrieben sind. Bei Letzteren entfällt die Trennung zwischen Vorklinik und Klinik und daher kann die Einteilung nicht vorgenommen werden. Es nahmen Studenten aller medizinischen Fakultäten Deutschlands teil, mit Ausnahme der privaten Medizinischen Hochschule Brandenburg (siehe Tabelle 9 in Appendix A). Tabelle 2 gibt einen Überblick über die weiteren demografischen Merkmale der Probanden. Da es aufgrund verschiedener Datenformate und -zeiträume nicht möglich war, die Daten aller statistischen Landesämter zu aggregieren, wurde stattdessen das „Berufsmonitoring Medizinstudierende“ der KBV in seiner dritten Befragungswelle von 2018 als Vergleich verwendet. Mit einer Stichprobengröße von 13915 Teilnehmern deckt sie die Gesamtpopulation mit einer hohen Genauigkeit ab und wird auch in anderen Studien als Vergleich herangezogen (z.B. Kasch et al., 2016). Alternativ wird aufgrund der Abwesenheit von Daten speziell für Medizinstudenten, die 21. Sozialerhebung des Deutschen Studentenwerkes, durchgeführt vom Deutschen Zentrum für Hochschul- und Wissenschaftsforschung, verwendet. Für das Merkmal Wohnort konnte keine Statistik speziell für Medizinstudenten oder Studenten allgemein gefunden werden. Ein Vergleich mit dem bundesdeutschen Durchschnitt erschien als nicht zielführend, da die meisten Universitäten in Mittel- und Großstädten liegen, so dass eventuell falsche Schlüsse über die Verteilung der Stichprobe gezogen werden würden.

3.4. Ökonometrische Modellierung

Im Rahmen dieser Arbeit wird eine Multinomial Logit Modellierung (MNL) auf Basis von McFadden (1974) verwendet. Eine genaue Beschreibung dieser Modellierung sowie ein Vergleich mit anderen Modellierungen kann bei Ryan et al. (2008) gefunden werden. Mithilfe des Random Utility Frameworks werden die individuellen Wahlpräferenzen ausgewertet. Dafür wird angenommen, dass die Probanden ihren Nutzwert („Utility“) maximieren wollen. Weiterhin wird angenommen, dass die Probanden ein komplettes Präferenzranking für die Verpflichtungsprogramme auf Basis dieser Nutzwertmaximierung besitzen. Daher werden sie immer dasjenige Verpflichtungsprogramm wählen, welches ihnen den höchsten Nutzwert verspricht, da es am besten zu ihren Präferenzen passt. Die Präferenzen werden durch die individuellen Merkmale der Probanden, wie Geschlecht, Herkunftsort, etc. beeinflusst. Die weitergehende Analyse basiert daher auf der folgenden Nutzwertfunktion nach Schlereth, Skiera, and Schulz (2018):

$$u_{n,i} = v_{n,i} + \varepsilon_{n,i} \tag{1}$$

Dabei stellt $u_{n,i}$ den Nutzwert von Probanden n während einer Teilnahme am Verpflichtungsprogramm i (Schlereth et al., 2018). Der Term $v_{n,i}$ stellt wiederum den deterministischen Teil der Formel dar. Weiterhin enthält er sämtliche beobachtbare Informationen, die während der Choice Sets sichtbar gemacht werden, wie z.B. die Ausprägungen der verschiedenen Eigenschaften. Der Term $\varepsilon_{n,i}$ ist der Fehlterm und enthält jegliche nicht beobachtbare Informationen (Schlereth et al., 2018). Der deterministische Nutzwert keiner Teilnahme am Verpflichtungsprogramm wäre damit durch $v_{n,0}$ gegeben.

Um die Wahrscheinlichkeit zu bestimmen, dass ein Proband n an einem bestimmten Verpflichtungsprogramm teilnimmt, kann die untenstehende Formel nach Ryan et al. (2008) angewandt werden. Hierbei wird von Homoskedastizität ausgegangen, so dass kein Faktor zum Skalieren des deterministischen Teils des Nutzwerts $v_{n,i}$ verwendet werden muss (Ryan et al., 2008).

Table 2: Deskriptive Statistik

Quelle: Eigene Erhebung, †KBV (2018a), ‡Middendorff et al. (2017a)

	Unabhängige Variable	Verteilung innerhalb der Erhebung (n = 335)	Vergleich
Studienabschnitt	Vorklinik	116 (34,6%)	31,7%†
	Klinik	173 (51,6%)	52,8%†
	Praktisches Jahr	33 (9,9%)	15,5%†
	Sonstiges	13 (3,9%)	k.A.†
Geschlecht	Weiblich	222 (66,3%)	65,6%†
	Männlich	112 (33,4%)	34,4%†
	Divers	1 (0,3%)	0%†
Alter	< 18 Jahre	2 (0,6%)	Bis 19 J.: 9,3%‡
	18-21 Jahre	79 (23,6%)	20-22 J.: 28,3%‡
	22-25 Jahre	143 (42,7%)	23-25 J.: 28,4%‡
	26-29 Jahre	77 (23%)	26-30 J.: 23,0%‡
	> 29 Jahre	34 (10,1%)	Über 31 J.: 10,2%‡
Herkunftsort	Großstadt (> 100000 Einwohner)	89 (26,6%)	45,8%†
	Mittelstadt (50000 – 100000 Einwohner)	27 (8,1%)	10,4%†
	Kleinstadt (20000 – 50000 Einwohner)	63 (18,8%)	25,3%†
	Landstadt (5000 – 20000 Einwohner)	70 (20,9%)	9,1%†
	Dorf (< 5000 Einwohner)	86 (25,7%)	9,6%†
Elternteil als Arzt	Beide Elternteile	15 (4,5%)	23,7%†
	Ein Elternteil	47 (14%)	
	Kein Elternteil	273 (81,5%)	76,3%†
Wohnort	Großstadt (> 100000 Einwohner)	235 (70,1%)	
	Mittelstadt (50000 – 100000 Einwohner)	47 (14%)	
	Kleinstadt (20000 – 50000 Einwohner)	34 (10,1%)	
	Landstadt (5000 – 20000 Einwohner)	10 (3%)	
	Dorf (< 5000 Einwohner)	9 (2,7%)	
Beziehungsstatus	Single, ohne Kind	136 (40,6%)	45,5%‡
	Single, mit Kind	2 (0,6%)	0,5%‡
	In einer festen Beziehung, ohne Kind	165 (49,3%)	46,0%‡
	In einer festen Beziehung, mit Kind	5 (1,5%)	2,0%‡
	Verheiratet, ohne Kind	11 (3,3%)	2,8%‡
	Verheiratet, mit Kind	13 (3,9%)	3,2%‡
	Geschieden	3 (0,9%)	k.A. ‡
	Verwitwet	0 (0%)	k.A. ‡

$$\Pr_{n,a}(i) = \frac{\exp\left(v_{n,i}\right)}{\sum_{j \in C_a} \exp\left(v_{n,j}\right)} \quad (2)$$

Dabei ist $\Pr_{n,a}(i)$ die Wahrscheinlichkeit, dass Proband n das Verpflichtungsprogramm i in jedem Auswahlentscheidungsblock a wählt. In Abgrenzung dazu steht j für die anderen Alternativen. N steht hierbei für die gesamte Anzahl der Probanden, während A die gesamte Anzahl der Auswahlentscheidungsblöcke darstellt. Ferner ist C_a die Gesamtheit der Entscheidungsalternativen für alle Vergütungen. Es wird davon ausgegangen, dass alle Entscheidungen voneinander unabhängig sind (Süptitz & Schlereth, 2017; Train, 2009).

Für die Schätzung der MNL müssen folgende Annahmen getroffen werden. Erstens muss der deterministische, systematische Teil des Nutzwertes $v_{n,i}$ durch das Produkt des Aufbauvektors X_i („design vector“) und des Präferenzvektors β_n („preference vector“) bestimmbar sein. Hierbei beschreibt der Aufbauvektor die Eigenschaften des Verpflichtungsprogramms und der Präferenzvektor die namensgebenden Präferenzen bezüglich dieser Eigenschaften. Zweitens muss die Varianz der Fehlerterme konstant sein (Homoskedastizität). Bei der Schätzung werden die β-Parameter so gewählt, dass die nachfolgende Log-Likelihood-Funktion maximiert wird (Ryan et al., 2008).

$$LnL = \sum_{n=1}^{N} \sum_{i \in C_a} y_{i,n} \left(X_i * \beta_n - \ln \sum_{j \in C_a} \exp(X_j * \beta_n) \right)$$
$$(n \in N, \alpha \in A) \quad (3)$$

Die oben beschriebene Formel basiert z.T. auf der Wahrscheinlichkeitsformel (2), da die Bestimmung der Parameter auf der Wahlwahrscheinlichkeit der Alternativen basiert. Zusätzlich wird hier $y_{i,n}$ als Indikator für das Wahlverhalten verwendet, wobei dieser den Wert eins annimmt, wenn Proband$n$ das Verpflichtungsprogramm i wählt, und ansonsten null beträgt.

Weitergehend liegt die Annahme der Unabhängigkeit von irrelevanten Alternativen („independence of irrelevant alternatives“, kurz IIA) zugrunde, so dass davon ausgegangen werden kann, dass die β-Parameter so geschätzt werden, dass sie auch bei der Nutzung einer kleineren Zahl von Entscheidungsalternativen konsistent sind (McFadden, 1974).

Der sogenannte Willingness-to-Accept-Wert beschreibt den minimalen Wert, bei dem ein Proband das Angebot der Verpflichtung annehmen würde. Da in dieser Arbeit untersucht wird, inwiefern Geld an den Medizinstudenten gezahlt werden muss, wird hier von einer Willingness-to-Accept gesprochen im Gegensatz zu einer Willingness-to-Pay wie in anderer Literatur (Schlereth et al., 2012). Die Berechnung verläuft jedoch analog und benutzt dabei die folgende Formel (Schlereth et al., 2012).

$$WTA_{n,i} = \frac{X_i * \beta_n}{\omega_n} \quad (4)$$

Wie vorhergehend beschreibt X_i den Aufbauvektor und β_n den Präferenzvektor. ω_n ist der Preisparamter. Auf diese Art und Weise kann bestimmt werden, wann ein Proband zwischen Annahme und Ablehnung des Verpflichtungsangebots indifferent wird. Die Implikationen der Verwendung einer „umgekehrten“ Willingness-to-Pay werden an späterer Stelle diskutiert.

4. Ergebnisse

In diesem Kapitel werden die Ergebnisse der Umfrage ausgewertet. Diese umfassen neben dem DCE auch die qualitativen Freitextfragen. Zuerst werden die Parameterschätzungen und Importance Weights ausgewertet und im zweiten Schritt die daraus ableitbare Willingness-to-Accept (WTA) und Akzeptanzwahrscheinlichkeiten berechnet. Im letzten Unterkapitel werden die vorhergehenden Ergebnisse mit der Analyse der qualitativen Umfrageteile verknüpft. Wo relevante Ergebnisse vorhanden sind, werden neben der Gesamtstichprobe auch Differenzierungen anhand demografischer oder anderer Merkmale getroffen.

4.1. Parameterschätzungen

Nachfolgend werden die Parameterschätzungen für die Eigenschaften und Ausprägungen, wie in Kapitel 3 beschrieben, präsentiert. Die Parameter wurden mit einer Random Utility Funktion, basierend auf einer lognormalen Verteilung, simuliert. Die Schätzung wurde mithilfe eines Hierarchical Bayer-Estimator der Software R durchgeführt. Somit stellen die β-Parameter die Wahrscheinlichkeit dar, dass ein durchschnittlicher Proband eine bestimmte Ausprägung einer Eigenschaft anstelle einer anderen wählt. Für den „Range of Level“-Effekt wurde durch die Auswahl der Eigenschaft kontrolliert. Die Eigenschaften wurden so gewählt, dass sie realistisch sind; insbesondere in Bezug auf die Breite des Preisparameters. Zusätzlich wurden sie so gewählt, dass sie die „Korrelationen zwischen den Eigenschaften auf null reduzieren“ (Green & Srinivasan, 1978). Außerdem wurden auch die Importance Weights für die Eigenschaften berechnet. Diese basieren auf der Spannweite der Parameterschätzungen, die untereinander ins Verhältnis gesetzt werden. Das Importance Weight beschreibt die Wichtigkeit einer Eigenschaft bei den im DCE getroffenen Entscheidungen. Somit weist ein hohes Importance Weight eine Eigenschaft aus, die einen großen Einfluss auf die Entscheidung hat.

Die nachfolgende Tabelle 3 zeigt die Parameter für die Gesamtstichprobe, die eine Größe von 335 Probanden besitzt. Zusätzlich zu den Schätzungen der β-Parameter werden auch deren Standardabweichung und die Importance Weights der Eigenschaften dargestellt.

Die β-Parameter entsprechen teilweise den Erwartungen. Die Teilnahme am Verpflichtungsprogramm selber, repräsentiert durch die Konstante, zeigt ein negatives Vorzeichen. Dies zeigt, dass eine Verpflichtung generell den Nutzwert

Table 3: Parameterschätzungen, Standardabweichung und Importance Weights für die Gesamtstichprobe (n = 335)

Quelle: Eigene Berechnungen

Eigenschaft	Ausprägung	β-Parameter	σ	Importance Weight
Konstante		-0,08	0,23	
(I) Ort der ersten Anstellung nach dem Studium	Landstadt (ca. 5000 Einwohner), ca. 30 min zum nächsten Ballungsgebiet	0,05	0,18	17,37%
	Landstadt (ca. 10000 Einwohner), ca. 60 min zum nächsten Ballungsgebiet	0,06	0,17	
	Dorf (ca. 1000 Einwohner), ca. 30 min zum nächsten Ballungsgebiet	0,18	0,19	
	Dorf (ca. 500 Einwohner), ca. 60 min zum nächsten Ballungsgebiet	-0,30	0,30	
(II) Monatliche Nettovergütung während des Studiums	300 € – 2000€	-0,0048	0,0235	25,04%
(III) Flexibilität der Arbeit	Vollzeit	-0,11	0,25	11,07%
	Teilzeit möglich	0,11	0,25	
(IV) Arbeitsstätte	Angestellt in einem Krankenhaus	0,04	0,18	11,35%
	Angestellt in einem MVZ	-0,01	0,18	
	Angestellt in einer Praxis	-0,03	0,24	
(V) Unterstützung bei der Facharztweiterbildung	Keine Unterstützung bei der Facharztweiterbildung	-0,17	0,18	13,77%
	Administrative Unterstützung bei der Facharztweiterbildung (z.B. Rotationsprogramm)	0,02	0,17	
	Administrative und inhaltliche Unterstützung bei der Facharztweiterbildung (z.B. Rotationsprogramm und Mentoring)	0,16	0,24	
(VI) Länge der Verpflichtung	1 Jahr	0,30	0,23	21,40%
	2 Jahre	0,14	0,16	
	3 Jahre	-0,45	0,30	

des durchschnittlichen Probanden reduziert, was den Erwartungen entspricht, da somit die Wahlfreiheit des Probanden eingeschränkt wird. Bei der ersten Eigenschaft zeigt die Ausprägung „Dorf (ca. 1000 Einwohner), ca. 30 min zum nächsten Ballungsgebiet“ den höchsten Nutzwert für den durchschnittlichen Probanden. Jedoch sieht man in der späteren Analyse, dass der Nutzwert für den jeweiligen Ort (d.h. der Ausprägung der ersten Eigenschaft) sehr heterogen innerhalb der Stichprobe verteilt ist. Der Preisparameter wurde als leicht negativ geschätzt. Der geringe Betrag war aufgrund der hohen Spanne von 300 € bis 2000 € zu erwarten. Ein negativer Preisparameter war zumindest bei einigen Probanden zu erwarten, da diese bei ihrem geplanten Praktizieren als Arzt vorzugsweise nichtmonetäre Ziele verfolgen. Außerdem wirkt sich die angebotene No-Choice-Option auf diesen Parameter aus, weil dadurch die Schätzung erschwert wird. Nichtsdestotrotz wurde der negative Durchschnittswert nicht erwartet. Weiteres dazu wird an späterer Stelle diskutiert. Ähnlich zur ersten Eigenschaft zeigt sich jedoch auch bei dem Preisparameter eine hohe Heterogenität. Für (III) Flexibilität der Arbeit zeigt sich die Parameterschätzung wie erwartet positiv für die Möglichkeit, Teilzeit zu arbeiten. Bezüglich der Arbeitsstätte gibt die Anstellung im Krankenhaus dem durchschnittlichen Probanden den höchsten Nutzwert, während eine Anstellung in einem MVZ oder einer Praxis gegenüber dem Krankenhaus weniger präferiert sind. Dies war aufgrund der Freitextantworten in der Vorstudie auch zu erwarten. Die Präferenz bezüglich der Arbeitsstätte zeigt auch eine Heterogenität innerhalb der Gesamtstichprobe. Bei der fünften Eigenschaft „Unterstützung bei der Facharztweiterbildung“ wurden die β-Parameter entlang der Erwartungen geschätzt. Mehr Unterstützung, d.h. administrative und inhaltliche Unterstützung, geben dem durchschnittlichen Probanden mehr Nutzwert. Die Schwankungen innerhalb der Gesamtstichprobe betreffen bei dieser Eigenschaft nicht das Vorzeichen. Lediglich das Ausmaß, wie viel Nutzwert aus der Unterstützung gezogen wird, variiert. Die Länge der Verpflichtung

wurde auch wie erwartet geschätzt. So gibt eine höhere Länge weniger Nutzwert, d.h. Probanden bevorzugen eine kürzere Länge. Auch diese Relation ist innerhalb der Stichprobe konsistent. Ähnlich zur vorhergehenden Eigenschaft schwankt jedoch das Ausmaß.

Bei den Importance Weights zeigt zwar der Preisparameter das höchste Gewicht, jedoch mit keinem großen Abstand zu den anderen Eigenschaften. Die Länge der Verpflichtung und der Ort der ersten Anstellung nach dem Studium stellen die zweit- bzw. drittwichtigsten Eigenschaften dar. Dies liegt innerhalb der Erwartung, da diese drei Eigenschaften die Eckpunkte des Verpflichtungsprogramms darstellen. Die Möglichkeit, Teilzeit zu arbeiten, stellt die am wenigsten wichtige Eigenschaft dar.

Die Präferenzen für den Ort, wie diese in der ersten Eigenschaft untersucht werden, sind heterogen. Im Rahmen der vier Ausprägungen der ersten Eigenschaft zeigen sich sowohl das Merkmal "Wunschort" als auch "Herkunftsort" und "Aktueller Wohnort" als geeignet, um zu differenzieren. Die Tabelle 4 zeigt die Parameterschätzungen für (I) Ort der ersten Anstellung nach dem Studium auf den angegebenen Wunschort heruntergebrochen. Hierbei wird sichtbar, dass sich abhängig vom Wunschort die Präferenzen für den Ort unterscheiden. Diejenigen Probanden, die in ein Dorf oder eine Kleinstadt ziehen möchten, haben weniger extreme Meinungen bezüglich der ersten Eigenschaft. Auch ist das Importance Weight für (I) Ort der ersten Anstellung nach dem Studium mit 11,02% geringer für den Wunschort Dorf und steigt auf 20,29% für den Wunschort Großstadt.

Für die Merkmale Herkunft sowie Wohnort zeigen sich ähnliche Zusammenhänge, welche weniger stark ausgeprägt sind. Jedoch herrscht dort ein weniger lineares Bild vor, da für die Mittelstadt die Parameterschätzung deutlich mehr zugunsten der letzten beiden Ausprägungen „Dorf“ ausfällt.

Somit lässt sich insbesondere für den Ort der ersten Anstellung nach dem Studium ein Verhalten mit Extrempunkten abbilden. Es gibt Probandengruppen, die einer Anstellung auf dem Land sehr negativ gegenüberstehen, und bei diesem DCE speziell eine Anstellung auf dem Dorf, welches weit vom Ballungsgebiet entfernt ist, stark ablehnen. Das hohe Importance Weight zeigt, dass diese extreme Meinung auch mit einer hohen Wichtigkeit verbunden ist, d.h. es ist ihnen wichtig, nicht in einem Dorf zu arbeiten. Andererseits gibt es andere Probandengruppen, die weniger negativ zur ländlichen Arbeit stehen. In Verbindung mit dem geringeren Importance Weight lässt es sich so lesen, als dass eine Anstellung auf dem Land für diese tolerabler ist–in Verbindung mit einem attraktiven Angebot im Rahmen des Verpflichtungsprogramms.

Der Preisparameter soll innerhalb des nächsten Unterkapitels im Rahmen der WTA beschrieben und diskutiert werden. Der β-Parameter für die (II) Monatliche Nettovergütung während des Studiums wird anders als die übrigen Parameter berechnet. Der Wert hängt, im Gegensatz zu den Präferenzen bei den anderen Parametern, mit dem Punkt zusammen, bei dem der Proband zwischen den Angeboten gleichgültig wird. Daher ist eine alleinige Interpretation des nummerischen Preisparameters ungenau. Zusätzlich hatten die Probanden die Möglichkeit, eine No-Choice-Option zu wählen, was die Berechnung weiterhin erschwert. Dies wirkt sich auch auf die Schätzung des Preisparameters aus, da sich einige Probanden unter keinen Umständen verpflichten würden und daher einen negativen Preisparameter haben. Eine große Besonderheit des Preisparameters ist das hohe Importance Weight, was die Wichtigkeit dieser Eigenschaft beschreibt, jedoch ist der Wert des Parameters nahe null. Dies kann so gesehen werden, als dass die Vergütung zwar wichtig für Probanden ist, jedoch im Durchschnitt keine extremen Meinungen dazu vorhanden sind. Ähnlich zum Ortsparameter zeigen sich auch hier die extremen Meinungen nur in Einzelbetrachtungen.

Die dritte Eigenschaft, welche die Möglichkeit, Teilzeit zu arbeiten, umfasst, gibt Probanden mit Kind einen wesentlich höheren Nutzwert. So liegt die durchschnittliche Parameterschätzung für Probanden mit Kind (n = 20) bei 0,21, während für ihre Kommilitonen ohne Kind (n = 312) bei 0,10 liegt[4]. Ähnliche, wenn auch nicht ganz so deutliche Trends zeigen sich in Bezug auf die Altersgruppe und den Beziehungsstatus, wobei ältere oder verheiratete Probanden einen höheren Nutzwert aus der Möglichkeit, Teilzeit zu arbeiten, ziehen. Die Importance Weights unterscheiden sich jedoch kaum.

Die Arbeitsstätte, entweder eine Anstellung in einem Krankenhaus, in einem MVZ oder einer Praxis, wurde im Rahmen der vierten Eigenschaft untersucht. Es ergaben sich geschlechtsspezifische Unterschiede bezüglich des Ausmaßes, bei denen weibliche Probanden die Anstellung in einem Krankenhaus (Parameterschätzung 0,06) gegenüber der Anstellung in einer Praxis (-0,05) bevorzugen, während sie zu einem MVZ leicht negativ (-0,01) eingestellt waren. Für ihre männlichen Kommilitonen lag auch eine leichte Präferenz zu Krankenhäusern (0,02) und gegen MVZs (-0,02) und Praxen (-0,00) vor. Ein weiterer Zusammenhang hierbei ist das Wunschfach. So passen die Präferenzen bezüglich der Arbeitsstätte auch zum Wunschfach. Die 46 Probanden, die Allgemeinmedizin als Wunschfach angegeben haben, bevorzugen die Arbeit in MVZs (0,02) und Praxen (0,01) gegenüber der Anstellung im Krankenhaus (-0,04). Dies passt zum Bild des Allgemeinmediziners, welcher hauptsächlich in MVZs oder Praxen tätig ist. Andersherum liegen für andere Wunschfächer, wie z.B. Chirurgie, klare Präferenzen für Krankenhäuser vor.

Das beispielhafte Verpflichtungsprogramm, auf dem das DCE basiert, hat außerdem eine Eigenschaft, bei der Unterstützung bei der Facharztweiterbildung angeboten wird. Generell zeigt die Gesamtstichprobe einen positiven Nutzwert davon, Unterstützung zu erhalten. Wenn zusätzlich zur administrativen Unterstützung auch noch inhaltliche Unterstützung durch z.B. Mentoring angeboten wird, wird der größte Nutzwert daraus gezogen. Dies gilt insbesondere für die 33 Probanden im PJ, für welche der entsprechende

[4] Drei Probanden gaben an geschieden zu sein. Für den Beziehungsstatus „Geschieden“ wurde nicht erfasst, ob der Proband ein Kind hat.

Table 4: Parameterschätzungen für die erste Eigenschaft nach Wunschort

Quelle: Eigene Erhebung

Ausprägung	Dorf	Landstadt	Kleinstadt	Mittelstadt	Großstadt
	β (σ)	β (σ)	β (σ)	β (σ)	β (σ)
Landstadt (ca. 5000 Einwohner), ca. 30 min zum nächsten Ballungsgebiet	-0,01 (0,14)	-0,01 (0,14)	0,07 (0,19)	0,04 (0,17)	0,07 (0,19)
Landstadt (ca. 10000 Einwohner, ca. 60 min zum nächsten Ballungsgebiet	0,05 (0,18)	0,11 (0,12)	0,08 (0,17)	0,06 (0,16)	0,02 (0,18)
Dorf (ca. 1000 Einwohner), ca. 30 min zum nächsten Ballungsgebiet	0,17 (0,16)	0,16 (0,21)	0,17 (0,17)	0,14 (0,18)	0,18 (0,17)
Dorf (ca. 500 Einwohner), ca. 60 min zum nächsten Ballungsgebiet	-0,21 (0,36)	-0,26 (0,24)	-0,32 (0,26)	-0,24 (0,3)	-0,28 (0,31)

Parameter im Durchschnitt auf 0,23 geschätzt wurde. Im Vergleich dazu haben die Probanden in der Vorklinik bzw. der Klinik durchschnittliche Parameterschätzungen von 0,16 bzw. 0,15. Auch ist das Importance Weight mit 17,61% für die Probanden im PJ höher als die 13,89% bzw. 12,97% in der Vorklinik bzw. Klinik und somit ist die Unterstützung für diese wichtiger.

Die letzte Eigenschaft, die Länge der Verpflichtung, verhält sich ähnlich eindeutig wie die Unterstützung bei der Facharztweiterbildung. Es ist eine generelle Präferenz zu einer kürzeren Verpflichtung zu erkennen. Zusätzlich treten kleinere Unterschiede im Ausmaß der Parameterschätzungen auf. So bevorzugen insbesondere Probanden in späteren Studienabschnitten die kürzeren Verpflichtungsdauern mehr als ihre Kommilitonen in früheren Abschnitten. Dies kann daran liegen, dass die Verpflichtung gegen Ende des Studiums einen konkreteren Eingriff in die Planungsfreiheit darstellt.

Alle Parameterschätzungen für die einzelnen Merkmale der Probanden, die im Rahmen dieses Unterkapitels beschrieben wurden, können in tabellarischer Form in Appendix B eingesehen werden.

4.2. Willingness-to-Accept

Im Rahmen dieses Unterkapitels werden anhand der Parameterschätzungen Berechnungen zu der Willingness-to-Accept (WTA) und den Akzeptanzwahrscheinlichkeiten verschiedener Programme vorgenommen. Die Berechnung der durchschnittlichen WTA unterliegt aufgrund der wechselnden Vorzeichen der Parameter abhängig vom Probanden einer leichten Verzerrung. So existieren Fälle eines negativen Preisparameters, was heißen würde, dass diese Probanden auch Geld für Verpflichtung zahlen würden. Außerdem gibt es einige Probanden, denen die Verpflichtung auch ohne die Zahlung einer Vergütung einen positiven Nutzwert gibt, obwohl sich diese in ihrer zukünftigen Laufbahn einschränken. Daher wurden für diese beiden Fälle, bei denen eine negative WTA berechnet werden würde, die WTA gleich null gesetzt. Dadurch zeigt sich im Durchschnitt eine leichte Verzerrung nach oben. Aus dem gleichen Grund ist die Betrachtung der WTA einzelner Ausprägungen weniger zielführend als die WTA ganzer Verpflichtungsprogramme. Genauere Implikationen werden in Kapitel 6 im Rahmen der Limitationen dieser Arbeit behandelt.

Die KV Brandenburg bietet zusammen mit dem Land Brandenburg ein Landärzteverpflichtungsprogramm an. Das Programm besteht seit 2019 und soll jährlich 25 Medizinstudenten mit bis zu 1000 € monatlich fördern, woraufhin diese sich verpflichten, nach Abschluss des Studiums für fünf Jahre in ländlichen Regionen in Brandenburg zu praktizieren (Landärztinnen/Landärzte-Richtlinie, 2019). Als ländliche Gebiete werden von der KV Brandenburg Klein- und Landstädte in Brandenburg, welche nicht direkt im Umland von Berlin liegen, wie z.B. Perleberg, Angermünde oder Lübbenau/Spreewald, definiert (KV Brandenburg, 2019). Dabei ist es den Verpflichtenden nicht vorgeschrieben, ob sie in einem Krankenhaus, einem MVZ oder als Angestellte eines Vertragsarztes praktizieren. Weiterhin ist die Förderung auf zehn Fachrichtungen, darunter Allgemeinmedizin sowie Kinder- und Jugendmedizin, limitiert. Von diesen wird ausgegangen, dass sie besonders in den zu fördernden ländlichen Gebieten benötigt werden (KV Brandenburg, 2019).

Anhand dieser Maßgaben kann die WTA dieses Programms für den durchschnittlichen Probanden der Umfrage berechnet werden. Aus den oben beschrieben Rahmenbedingungen lassen sich die Eigenschaften wie folgt in ihren Ausprägungen für die Berechnung wählen. Für (I) Ort der ersten Anstellung nach dem Studium wird die zweite Ausprägung „Landstadt (ca. 10000 Einwohner), ca. 60 min zum nächsten Ballungsgebiet“ verwendet. Da bezüglich (III) Flexibilität der Arbeit und (V) Unterstützung bei der Facharztweiterbildung keine weiteren Angaben gemacht werden, wird hier von jeweils der ersten Ausprägung ausgegangen. Da die Arbeitsstätte nicht spezifiziert wird, werden die Wahrscheinlichkeiten für alle drei Ausprägungen von (IV) Arbeitsstätte berechnet. Für (VI) Länge der Verpflichtung wird die Ausprägung „3 Jahre“ gewählt, da im Rahmen der Umfrage die Option von fünf Jahren nicht inkludiert wurde. An späterer

Stelle in diesem Unterkapitel wird kurz beschrieben, wie die Attraktivität des Programms durch Anpassungen erhöht werden kann. Tabelle 5 zeigt die WTAs für das so modifizierte Verpflichtungsprogramm der KV Brandenburg differenziert nach Arbeitsstätte.

Für dieses leicht modifizierte, d.h. auf drei Jahre verkürzte Verpflichtungsprogramm liegt die durchschnittliche WTA bei 1469,55 € im Monat. Je nach Arbeitsstelle schwankt die WTA zwischen 1525,24 € und 1412,78 € im Monat, wobei sie für die Arbeit im MVZ am höchsten und für die Arbeit in der Praxis am niedrigsten ist. Damit liegt die WTA zwischen 110 € und 390 € höher als das Gehalt einer Krankenschwester in Ausbildung im Monat (verdi., 2018) bzw. zwischen 310 € und 630 € für einen Medizintechniker in Ausbildung im Monat. Somit beträgt die WTA deutlich mehr für das Medizinstudium als für einen Ausbildungsberuf im gleichen Sektor. Außerdem ist die Höhe der WTA deutlich über den durchschnittlichen Einkommen eines Studenten in Deutschlands (Middendorff et al., 2017b), jedoch noch unter den durchschnittlichen Konsumausgaben eines deutschen Einpersonenhaushaltes mit etwas über 1700 € /Monat (Statistisches Bundesamt, 2020). Damit befindet sich die durchschnittliche WTA in einem realistischen Preisband. Ein Haushaltseffekt kann beobachtet werden, wenn man indirekt mit dem Beziehungsstatus für die (zukünftige) Haushaltsgröße kontrolliert. Die unten gezeigte Tabelle 5 bietet einen Überblick über die WTAs für das gleiche Programm nach Beziehungsstatus.

Die WTA verläuft damit z.T. kongruent mit den nach Haushalt benötigten Ausgaben, sowie der Erwartung zukünftiger Ausgaben. Für den alleinstehenden Medizinstudenten ist die WTA am niedrigsten; dieser wohnt am wahrscheinlichsten in einer Wohngemeinschaft, einemWohnheim oder bei den Eltern. Folglich sind für diesen die durchschnittlichen Kosten auch am geringsten (Middendorff et al., 2017b). Sobald sich der Student in einer Beziehung befindet, ist es unwahrscheinlicher, dass er in einer dieser Wohnformen verbleibt, so dass die Kosten des Wohnens, aber auch andere Ausgaben steigen (Statistisches Bundesamt, 2020). Zusätzlich kann die geringere WTA für verheiratete Probanden dadurch erklärt werden, dass die Ehe vermögensrechtliche Auswirkungen hat und auf das Einkommen des Partners gezählt werden kann.

Zu beachten ist jedoch, dass es sich um die durchschnittliche WTA handelt und in den Einzelbetrachtungen Extrempunkte vorliegen können. Ähnlich wie bei den Parameterschätzungen zeigen sich verschiedene Probandengruppen. So existieren diejenigen, die eine sehr hohe WTA besitzen. Es ist anzunehmen, dass diesen ihre Freiheit wichtig ist und durch die Verpflichtung eine Einschränkung vorliegt. Andererseits gibt es eine breite Gruppe an Probanden, die eine sehr niedrige WTA oder eine WTA gleich null haben und sich daher „umsonst“ verpflichten würden.

Alternativ können die Akzeptanzwahrscheinlichkeiten für die verschiedene Arbeitsstätten berechnet werden. Hierbei wird für (II) Monatliche Nettovergütung während des Studiums von 1000 € ausgegangen. Tabelle 7 zeigt die Akzeptanzwahrscheinlichkeiten für die drei möglichen Arbeitsstätten unter Berücksichtigung des Studienabschnittes.

Die Akzeptanzwahrscheinlichkeit von durchschnittlich 44,0% für das auf drei Jahre modifizierte Programm sinkt über den Verlauf des Studiums. Hierbei ist davon auszugehen, dass ein Treiber das Verhältnis der Verpflichtungszeit zur verbleibenden Studienzeit ist. So ist eine Verpflichtung von drei weiteren Jahren am Anfang der Studienzeit von über sechs Jahren akzeptabler, da die Zeit der Verpflichtung weiter in der Zukunft liegt und das Medizinstudium verhältnismäßig lange dauert. Wenn jedoch der Abschluss des Studiums bevorsteht, stellt eine Verpflichtung einen konkreteren Eingriff in die persönliche Planung dar. Ein ähnlicher, wenn auch nicht ganz so deutlicher Trend zeigt sich auch in Bezug auf die Akzeptanzwahrscheinlichkeiten basierend auf dem Beziehungsstatus. Dabei stellt jedoch, wie im Literaturüberblick erwähnt, die Abhängigkeit vom Partner einen wichtigen Treiber dar.

Allerdings kann durch Anpassungen des Programms die Akzeptanzwahrscheinlichkeit erhöht und die WTA gesenkt werden. Ein starker Effekt zeigt sich dabei, wenn das Programm um die Unterstützung bei der Facharztweiterbildung erweitert wird. Werden den Probanden administrative und inhaltliche Unterstützung geboten, steigt die Akzeptanzwahrscheinlichkeit von durchschnittlich 44,0% auf 49,6%. Noch deutlicher wird der Wert der Unterstützung für die Probanden bei der Betrachtung der WTA. Diese fällt von durchschnittlich 1469,55 € auf 774,19 € im Monat. Bei einem Unterschied in dieser Größenordnung sollte es in Betracht gezogen werden, das Programm der KV Brandenburg derartig erweitert werden sollte.

Ein weiteres Programm, das im Rahmen dieses Unterkapitels untersucht werden soll, ist das Stipendium im Rahmen der „Landarzt werden“-Initiative des Kreises Waldeck-Frankenberg in Nordhessen. Es unterscheidet sich in einigen Punkten deutlich von dem Programm der KV Brandenburg. Der Landkreis Waldeck-Frankenberg zahlt 400 € im Monat während des Studiums für eine fünfjährige Verpflichtung (Annäherung durch „3 Jahre“). Um das Programm attraktiver zu machen, hebt die Initiative jedoch deutlich die Familienfreundlichkeit heraus, indem Teilzeitarbeit aktiv angeboten wird und hauptsächlich für Praxen gesucht wird (Landkreis Waldeck-Frankenberg, 2017). Der Ort wird durch die dritte Ausprägung der ersten Eigenschaft „Dorf (ca. 1.000 Einwohner), ca. 30 min zum nächsten Ballungsgebiet“ angenähert. Für die fünfte Eigenschaft „Unterstützung bei der Facharztweiterbildung“ wird keine Unterstützung angeboten. Die Berechnung der WTA erfolgt ansonsten auf die gleiche Art und Weise wie beim Programm der KV Brandenburg. Hierbei liegt die durchschnittliche WTA für die untersuchte Preisspanne bei 1111,73 € im Monat. Diese ist deutlich höher als die vom Programm gebotenen 400 € monatlich. Weiterhin kann die Akzeptanzwahrscheinlichkeit für dieses Programm berechnet werden. Für die beschriebenen Ausprägungen beträgt diese im Durchschnitt 42,7%.

Jedoch gilt für eine theoretische Erfolgsbeurteilung anhand dieser Zahlen zu beachten, dass das Programm auf

Table 5: Willingness-to-Accept für das modifizierte Landärzteprogramms der KV Brandenburg

Quelle: Eigene Berechnung, Landärztinnen/Landärzte-Richtlinie (2019)

Arbeitsstätte	**Willingness-to-Accept Preisspanne 300 € - 2000 €**
Angestellt in einem Krankenhaus	1470,62 € pro Monat
Angestellt in einem MVZ	1525,24 € pro Monat
Angestellt in einer Praxis	1412,78 € pro Monat
Durchschnitt	**1469,55 € pro Monat**

Table 6: Willingness-to-Accept für das modifizierte Landärzteprogramms der KV Brandenburg nach Beziehungsstatus

Quelle: Eigene Erhebung

Arbeitsstätte	**Willingness-to-Accept Preisspanne 300€ - 2000€**		
	Single (n = 138)	**Feste Beziehung (n = 170)**	**Verheiratet (n = 24)**
Angestellt in einem Krankenhaus	823,64 € pro Monat	2056,50 € pro Monat	1214,03 € pro Monat
Angestellt in einem MVZ	896,97 € pro Monat	2101,93 € pro Monat	1232,45 € pro Monat
Angestellt in einer Praxis	971,18 € pro Monat	1843,41 € pro Monat	1069,36 € pro Monat
Durchschnitt	**897,26 € pro Monat**	**2000,61 € pro Monat**	**1171,95 € pro Monat**

Table 7: Akzeptanzwahrscheinlichkeiten für das modifizierte Landärzteprogramm der KV Brandenburg

Quelle: Eigene Berechnung, Landärztinnen/Landärzte-Richtlinie (2019)

Arbeitsstätte	**Akzeptanzwahrscheinlichkeit**			
	Vorklinik (n = 116)	**Klinik (n = 173)**	**PJ (n = 33)**	**Durchschnitt (n = 335)**
Angestellt in einem Krankenhaus	48,5%	45,2%	33,1%	**44,7%**
Angestellt in einem MVZ	48,2%	43,7%	33,5%	**43,9%**
Angestellt in einer Praxis	47,7%	43,5%	31,6%	**43,4%**
Durchschnitt	**48,1%**	**44,1%**	**32,7%**	**44,0%**

jeweils drei Studenten pro Jahr über die Dauer von drei Jahren ausgelegt ist. So finden sich allein in der Gesamtstichprobe (n = 335) insgesamt 70 Probanden mit einer WTA größer 0, aber kleiner oder gleich den 400 € im Monat, die das Programm bietet. Von diesen 70 Probanden haben alle eine Akzeptanzwahrscheinlichkeit, welche über 50% liegt, so dass basierend auf diesen Zahlen und dieser Probandengruppe der Landkreis Waldeck-Frankenberg sein Programm problemlos füllen könnte. Ähnlich wie zum Programm der KV Brandenburg könnte auch der Landkreis Waldeck-Frankenberg die Attraktivität seines Programms durch Unterstützung bei der Weiterbildung steigern. Auch wenn die inhaltliche Unterstützung die Mittel des Landkreises überstiege, kann allein durch die Bereitstellung administrativer Unterstützung die WTA weit gesenkt werden und die Akzeptanzwahrscheinlichkeit gesteigert werden. Eine Rotation durch die für den Facharzt benötigten Stellen sollte innerhalb des Landkreises möglich sein und würde die WTA von 1111,73 € auf durchschnittlich 659,19 € im Monat reduzieren sowie die Akzeptanzwahrscheinlichkeit von 42,7% auf 46,8% anheben.

Für alle Berechnungen gilt jedoch zu beachten, dass die Eigenschaft (I) Ort der ersten Anstellung nach dem Studium nicht die vielschichtigen Effekte der Realität abdeckt. Die größte Limitation für den Fall der KV Brandenburg ist jedoch, dass das Land Brandenburg abgesehen von der privaten Medizinischen Hochschule Brandenburg, die weniger als 500 Studenten hat (Amt für Statistik Berlin-Brandenburg, 2020), keine medizinische Fakultät besitzt. Daher ist es fragwürdig, ob sowohl die WTA als auch die Akzeptanzwahrscheinlichkeiten realistisch sind, da potentielle Teilnehmer am Programm dafür nach Brandenburg umziehen müssten. Ähnliches gilt für das Programm des Landkreises Waldeck-Frankenberg. Außerdem wurde die Verpflichtungsdauer durch drei Jahre statt fünf angenähert, so dass in Realität die WTA höher bzw. die

Akzeptanzwahrscheinlichkeit niedriger sein wird. Weitere Einflüsse und Limitationen werden in Kapitel 6 diskutiert.

4.3. Qualitative Analyse

Dieses Unterkapitel dient der Analyse des qualitativen Teils der Umfrage sowie der Verknüpfung dieser Ergebnisse mit dem quantitativen Teil. Neben dem DCE hatten die Probanden die Möglichkeit, ihre Präferenzen direkt anzugeben. Zusätzlich konnten sie im Rahmen von zwei Freitextfragen nach dem DCE sowie in einem abschließenden Kommentarfeld weitere Anmerkungen geben. Von diesen Möglichkeiten haben 290 der 335 Probanden durch zumindest eine Freitextantwort Gebrauch gemacht.

Für die Einstellung gegenüber Verpflichtungsprogrammen zeigt sich ein leicht differenziertes Bild. Auf einer fünfstufigen Likertskala wurde die Zustimmung zur Aussage „Ich finde es gut, Studenten die Möglichkeit zu geben, sich für finanzielle Unterstützung während des Studiums zur Arbeit auf dem Land zu verpflichten." abgefragt. Hierbei gaben 56 Probanden (16,7%) an, dieser Aussage völlig zuzustimmen. 129 Probanden (38,5%) stimmten zu, 50 Probanden (14,9%) standen der Aussage neutral gegenüber, während 78 Probanden (23,3%) diese ablehnten und 22 Probanden (6,6%) diese völlig ablehnten. Somit besteht im Durchschnitt eine leichte Tendenz zur Zustimmung. Im Vergleich der demografischen Gruppen zeigten sich Probanden aus früheren Studienabschnitten sowie solche, die in Land- oder Kleinstädten aufgewachsen sind, etwas positiver gegenüber Verpflichtungsprogrammen.

Weiterhin lässt sich die Einstellung gegenüber Verpflichtungsprogrammen auch in den Zusammenhang mit den vorher berechneten WTAs setzen, wobei sich ein Zusammenhang in umgekehrter U-Form zeigt. Diejenigen Probanden, die einer Verpflichtung am positivsten gegenüberstehen, haben eine durchschnittliche WTA von 761,77 € für das KV Brandenburg Programm mit einer Stelle im Krankenhaus. Für die Probanden, die einer Verpflichtung neutral gegenüberstehen, steigt die WTA auf 4766,86 € an. Die Probandengruppe, die Verpflichtungsprogramme völlig ablehnt, hat hingegen eine durchschnittliche WTA von 404,52 € . An dieser Stelle lässt sich davon ausgehen, dass diese Zahl aufgrund der Modellierung und der Berechnung der WTA auftritt und nicht die Realität wiederspiegelt. Eine genauere Diskussion der Limitationen der Modellierung und der Berechnung findet sich in Kapitel 6. Um die WTAs in Verbindung mit der Einstellung gegenüber Verpflichtungen in ein besseres Verhältnis zu setzen, können zusätzlich noch die Akzeptanzwahrscheinlichkeiten analysiert werden. Diese sinken von 57,1% für die Probanden, die völlig zustimmen, auf 27,1% für die Probanden, die Verpflichtungsprogramme völlig ablehnen.

Bei den angegebenen Präferenzen der Wunschorte zeigt sich eine klare Präferenz zu größeren Städten. Jedoch gab zunächst etwas weniger als ein Drittel (97 Probanden, d.h. 29,0%) an, noch keinen Wunschort zu haben. Ferner sahen sich 11 Probanden (3,3%) später in einem Dorf, während 31 Probanden (9,3%) bzw. 40 Probanden (11,9%) Land- bzw. Kleinstädte bevorzugen. Weitere 56 Probanden (16,7%) bevorzugen Mittelstädte und 100 Probanden (29,8%) Großstädte. Der Wohnort zeigte aufgrund des hohen Vorkommens der Großstadt als aktueller Wohnort kaum nennenswerte Zusammenhänge mit dem Wunschort. Jedoch wurde ersichtlich, dass, abgesehen von der Mittelstadt, der Herkunftsort auch am häufigsten den Wunschort darstellt. Am deutlichsten wird dieser Zusammenhang für Dorf und Großstadt, wobei 81,8% bzw. 49,0% der Probanden, welche auf dem Dorf bzw. in der Großstadt aufgewachsen sind, diesen Ort wieder als Wunschort angeben. Zusätzlich lässt sich ein Trend zu den Extremen hin erkennen, da die Probanden aus einer Mittelstadt die gleichverteiltesten Präferenzen bezüglich des Wunschortes zeigen. In einem geringeren Ausmaß gilt dies auch für die Kleinstadt.

Außerdem gab in etwa ein Drittel (109 Probanden, d.h. 32,5%) an, noch keine Wunschfachrichtung zu haben. Die beliebteste Fachrichtung war die Allgemeinmedizin, die von 46 Probanden (13,7%) angegeben wurde. Diese wird dicht gefolgt von der Inneren Medizin mit 41 Probanden (12,2%) und der Chirurgie mit 35 Probanden (10,5%). Von den 34 möglichen Fachrichtungen wurden 13 kein einziges Mal gewählt. Tabelle 11 in Appendix A zeigt die genaue Verteilung. Weiterhin liegt in Bezug auf die Allgemein- und Innere Medizin keine deutliche Abweichung in den Präferenzen zwischen den Geschlechtern vor, während die Frauenheilkunde als deutlich weiblich, die Chirurgie hingegen als deutlich männlich besetztes Fachgebiet heraussticht.

Im Rahmen der ersten Freitextfrage hatten die Probanden die Möglichkeit, weitere Faktoren zu nennen, die für sie bei der Entscheidung, an einem Verpflichtungsprogramm teilzunehmen, wichtig wären. Dabei zielte diese Frage auf die Phase der *Rekrutierung* von Landärzten ab (siehe Kapitel 2). Diese wurden bei der Auswertung in 20 Kategorien eingeordnet, wobei 16 der Kategorien neu waren und die verbleibenden vier Kategorien die Eigenschaften (III) bis (VI) widerspiegeln. Die Kategorien wurden nach einer ersten Durchsicht der Freitextantworten erstellt und im Verlauf der Einordnung kontinuierlich angepasst, um eine genaue Abdeckung zu gewährleisten. Eine Übersicht und Beschreibung der 16 neuen Kategorien kann in Tabelle 16 in Appendix C gefunden werden. Zusätzlich werden beispielhafte Zitate der Freitextantworten zur Illustration verwendet. Für die Kategorie „Genauer Ort" ergeben sich logische Überschneidungen mit anderen Kategorien wie z.B. „Familiennähe". Diese werden auch beschrieben. Die folgende Abbildung 3 stellt die Häufigkeit der Nennungen aller 20 Kategorien dar, während Abbildung 4 aufzeigt, welche Kategorien zusammen genannt werden. Hierbei wird für die bessere Darstellung die Größe der Punkte, welche die Anzahl der Gemeinsamnennungen zeigt, logarithmisch gestaucht.

Die Spannweite der weiteren Faktoren, die für die Entscheidung am Verpflichtungsprogramm teilzunehmen als wichtig angesehen werden, ist weit und zeigt, ähnlich wie die Gesamtstichprobe, eine große Heterogenität auf. Mit Abstand am häufigsten (82 von 260 Antworten, d.h. 31,5%) tritt der genaue Ort als weiterer relevanter Faktor auf. Hier-

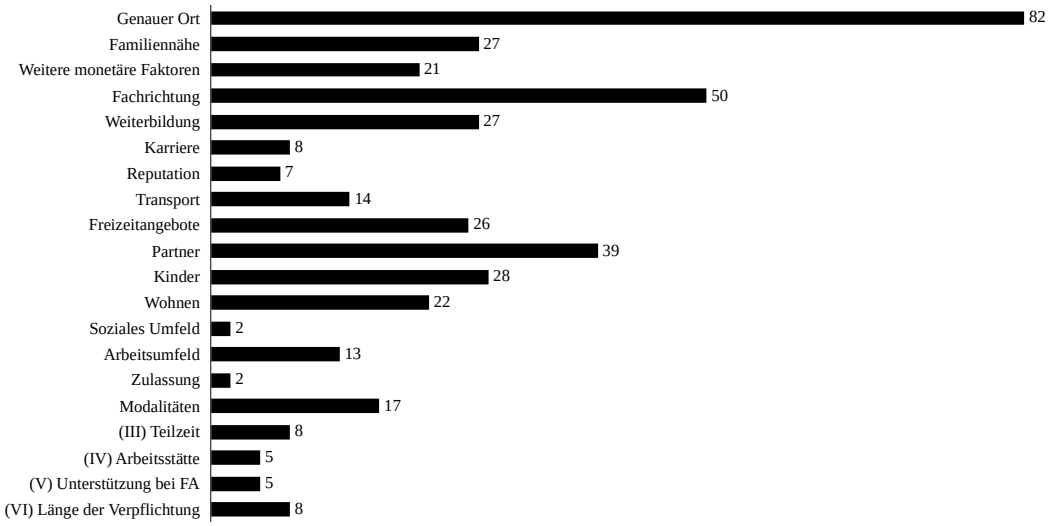


Figure 3: Qualitativ genannte, wichtige Faktoren für die Entscheidung der Verpflichtung (Mehrfachnennungen möglich)

Quelle: Eigene Erhebung

bei reicht die Spezifikation vom Bundesland über den Landkreis bis zur Gemeinde. In diesem Rahmen wurde auch der geografische Rahmen des Verpflichtungsprogramms innerhalb der Bundesrepublik angesprochen. So beschreibt Proband 286 beispielsweise die Situation wie folgt: „es müsste Bundesländerübergreifend sein (also nicht nur, wenn ich in [Bundesland 1] studiere, bekomme ich das Geld für eine Praxis in [Bundesland 1], sondern auch wenn ich in [Nachbarbundesland] studiere (ich bin nämlich gebürtig aus [Bundesland 1] und kann an diesem Programm nicht teilnehmen)". Häufig wird der genaue Ort in Verbindung mit der Kategorie „Familiennähe" angegeben, d.h. dass der Proband gerne die Verpflichtung an einem bestimmten Ort in Familiennähe absolvieren würde. Probanden, die den genauen Ort (in Familiennähe) bevorzugen, sind zu 73,0% (76,9%) weiblich, was deutlich über dem Stichprobendurchschnitt von 66,3% liegt. Zusätzlich sind sie, wenn sie in die Nähe der Familie ziehen wollen, eher beziehungsmäßig ungebunden mit 50,0% im Vergleich zu 41,2% für die Gesamtstichprobe.

Die zweithäufigste Kategorie ist die freie Auswahl der Fachrichtung, die in 19,2% der Antworten genannt wurde. Probanden, für die die freie Auswahl der Fachrichtung wichtig ist, gaben in der Umfrage als Wunschfach häufig Fachrichtungen an, die einen hohen Spezialisierungsgrad erfordern und für die Tätigkeit eines Landarztes als Hausarzt weniger benötigt werden. Dazu zählen die Arbeitsmedizin, Mund-Kiefer-Gesichtschirurgie, Neurochirurgie, Rechtsmedizin und Strahlentherapie. Darüber hinaus traten keine relevanten Abweichungen innerhalb der Stichprobe für andere Probandenmerkmale auf.

Die Kategorien „Partner" und „Kinder" zeigen zu erwartende Beziehungen zu den Merkmalen der Probanden. So werden Antworten aus den Kategorien „Partner" überwiegend von liierten (68,4% verglichen zu 50,7% der Gesamtstichprobe) und verheirateten (15,8% ggü. 7,2%) Probanden genannt. Gleiches gilt für Antworten aus der Kategorie „Kinder", welche in 28,6% von Probanden mit Kindern kamen (5,7% der Gesamtstichprobe).

Weiterhin treten andere Antworten überwiegend bei einem Geschlecht der Probanden auf. So ist für weibliche Probanden das Arbeitsumfeld deutlich wichtiger als für ihre männlichen Kommilitonen (76,9% im Vergleich zu 66,3% in der Gesamtstichprobe), während für männliche Probanden die Karriere und Reputation eine hohe Wichtigkeit hat (50,0% bzw. 71,4% vs. 33,4%). Auch nennen mehr Männer das Gehalt als weiteren wichtigen Faktor für die Entscheidung (60,0% vs. 33,4%).

Weitere Antworten zu anderen Kategorien treten gleichverteilter über die Stichprobe auf. So lassen sich für „Transport" sowie „Freizeitangebote" keine Zusammenhänge mit den Merkmalen der Probanden erkennen. Gleiches gilt auch für „Wohnen". Die Kategorien „Zulassung" und „Soziales Umfeld" wurden jeweils nur zweimal erwähnt, so dass eine Analyse hier nicht zielführend ist. Nichtsdestotrotz haben diese Antworten ihre Wichtigkeit und sollen in der Diskussion dieser Arbeit mit aufgegriffen werden.

Wichtig für das anschließende Kapitel sind die Antworten in der Kategorie „Modalitäten". Hier haben die Probanden ihre relevanten Faktoren für die Entscheidung der Teilnahme in Fragen formuliert. Ein Punkt, welcher sich abzeichnet, ist die Bewerbung für das Verpflichtungsprogramm und der damit verbundene Aufwand. Weiterhin ist auch der Zeitpunkt, zu dem die Verpflichtung eingegangen wird, eine wichtige Modalität. Außerdem wird häufig der Rücktritt von der bzw. Kündigung der Verpflichtung erwähnt. Dabei sind vor allem die Konsequenzen, d.h. Rückzahlung der Vergütung, weitere mögliche Vertragsstrafen, und eventuelle, mildernde Umstände, z.B. familiäre Gründe, wichtig.

Zusätzlich wurden bei der ersten Freitextfrage noch sieben Antworten gegeben, die keiner Kategorie zu geordnet

werden konnten. Für zwei Probanden wäre weitere Unterstützung für entweder Studienaufenthalte oder Praktika im Ausland wichtig. Jeweils einmal als weitere wichtige Faktoren genannt wurden „Ausländerfeindlichkeit des Ortes der Anstellung“, „Dienste“, „Einarbeitungszeit der Stelle“, „Erfahrungsberichte vorheriger Teilnehmer“ und „Seriosität des Programms“.

Wie in Abbildung 4 gezeigt, treten bei den weiteren relevanten Faktoren Themenkomplexe, nachfolgend Cluster genannt, auf, da einige Kategorien bevorzugt zusammen genannt werden.

Wie vorher umrissen, werden Wünsche nach einem genauen Ort oft präzisiert. Am häufigsten soll der genaue Ort in Familiennähe liegen (27 Nennungen). Auch soll der genaue Ort genug Freizeitmöglichkeiten bieten (11 Nennungen) und in der Nähe des Partners liegen (8 Nennungen). Ebenso wurde Aspekte der Kategorie „Wohnen“, wie die Verfügbarkeit von Wohnungen, in diesem Zusammenhang genannt. Jedoch liegt es in der Natur der Kategorie „Genauer Ort“ weiter zu präzisiert werden.

Das anderweitig häufigste Cluster betrifft die Gründung einer eigenen Familie und besteht aus Antworten aus den Kategorien „Partner“ und „Kinder“, welche zusammen in 18 Fällen auftreten. Entweder wurden diese von den Probanden einzeln genannt als „Partner“ und „Kinder“ oder direkt in dem Begriff „Familie“ verbunden. Dabei gab es Antworten wie „gute Optionen zu Familiengründung“ (Proband 131) oder „Lebensumstände wie Partner und/oder Kinder beziehungsweise deren Planung“ (Proband 336). Probanden, die Antworten aus den Kategorien „Partner“ und „Kinder“ gaben, sind vornehmlich weiblich (15 der 18 Nennungen)

Ein weiteres Cluster ist „fachbezogene Weiterbildung“, bestehend aus „Fachrichtung“ und „Weiterbildung“, mit 10 Nennungen. Hierbei könnten jedoch keine relevanten Zusammenhänge mit den Merkmalen der Probanden beobachtet werden.

Weitere Korrelationen können gefunden werden, allerdings ist ein inhaltlicher Zusammenhang zwischen den genannten Kategorien fragwürdig.

Die zweite Freitextfrage ging über den Zeitraum der Verpflichtung hinaus und bat die Probanden eine begründete Einschätzung abzugeben, ob diese über den Zeitraum der Verpflichtung hinaus auf dem Land bleiben würden. Dabei zielt diese Freitextfrage auf die *Erhaltung* von Landärzten und relevante Faktoren dafür ab (siehe Kapitel 2). Die Antworten wurden einmal nach Zustimmung in die Kategorien „Ja“, „Eher ja“, „Unentschieden“, „Eher nein“, „Nein“ eingeteilt. Sollte dies aus dem Text nicht nachvollziehbar sein, wurde die Antwort in die Kategorie „Unklar“ eingeteilt. Dies geschah in 13 Fällen. Abbildung 5 zeigt die Verteilung über die Zustimmung zur zweiten Freitextfrage.

Der Großteil der Probanden steht einer weiteren Zeit auf dem Land nach Ablauf der Verpflichtung neutral bis positiv gegenüber. Insbesondere die weiblichen Probanden sehen sich nach dem Ende der Verpflichtung weiterhin auf dem Land mit 52,7%, die „Ja“ oder „Eher ja“ geantwortet haben. Dem gegenüber haben nur 37,5% der männlichen Probanden positiv geantwortet. Umgekehrt ist der Anteil der männlichen Probanden in den Kategorien von „Unentschieden“ bis „Nein“ höher. Zusätzlich zeigt sich auch, dass verheiratete Probanden oder solche in festen Beziehungen auch eher auf dem Land bleiben würden, nachdem die Verpflichtung abgeleistet ist. Sie sagen in 57,1% bzw. 53,1% aus, dass sie nach Abschluss der Verpflichtung auf dem Land bleiben würden, während das nur 39,4% der Alleinstehenden tun.

Im nächsten Schritt der Analyse wurden dann die Zustimmungsantworten auf ihre Begründung analysiert. Hierbei wurden dieselben Kategorien wie bei der ersten Freitextantwort verwendet. Dadurch lassen sich dann folgende logische Konstruktionen bilden:

- „Ja/Eher ja, wenn gegeben [Kategorie]“
- „Unentschieden, abhängig von [Kategorie]“
- „Nein/Eher nein, aufgrund von [Kategorie]“

Somit lassen sich die inhaltlichen Kategorien mit den Zustimmungskategorien verbinden und dezidierte Gründe und Bedingungen für den Verbleib bzw. den Nichtverbleib können untersucht werden. Die nachfolgende Tabelle 8 zeigt eine Übersicht der Nennungen der inhaltlichen Kategorien in Verbindung mit den Zustimmungskategorien. Für die 13 Antworten, welche als „Unklar“ kategorisiert wurden, war eine Einteilung nicht möglich.

Die wichtigsten Faktoren für einen Verbleib auf dem Land sind das Arbeitsumfeld (28 Nennungen), Möglichkeiten und Zufriedenheit des Partners (20 Nennungen) und Transport (14 Nennungen). Das Arbeitsumfeld ist vor allem für die weiblichen Probanden besonders wichtig (21 von 27 Nennungen). Gleiches gilt auch für Partner und Kinder (18 von 20 bzw. 17 von 21 Nennungen). Hier wird häufig von Arbeitsplätzen für den Partner und Betreuungsangebote (z.B. Kindertagesstätte, Schule) für die Kinder gesprochen. Bei den männlichen Probanden sticht kein Erhaltungsfaktor stark heraus. Antworten aus der Kategorie „Soziales Umfeld“ thematisieren hauptsächlich die Integration des Landarztes und des Partners in die ländliche Gemeinschaft. Hierbei sprechen sie neben Familie und Freunde auch Punkte an, die eher für das Leben auf dem Land typisch sind, wie z.B. Sportvereine und Ortsfeste. Ansonsten werden an dieser Stelle auch die Freizeitangebote auf dem Land thematisiert. So heben mehrere Probanden die Natur und Möglichkeiten, an der „frischen Luft“ Sport zu machen, als Grund auf das Land zu gehen und dort zu bleiben hervor.

Für die unentschiedenen Probanden wird neben den oben aufgelisteten Gründen häufig ein abstraktes „Wenn es mir gefällt“ verwendet. Dieses konnte in keiner zielführenden Weise kategorisiert werden, wenn keine weitere Präzisierung getroffen wurde. Nichtsdestotrotz zeigt dies die generelle Unsicherheit in dieser Gruppe. Weitere Faktoren wie Arbeits- und soziales Umfeld sowie Partner und Kinder treten ähnlich wie bei der Gruppe, die auf dem Land bleiben würde, auf.

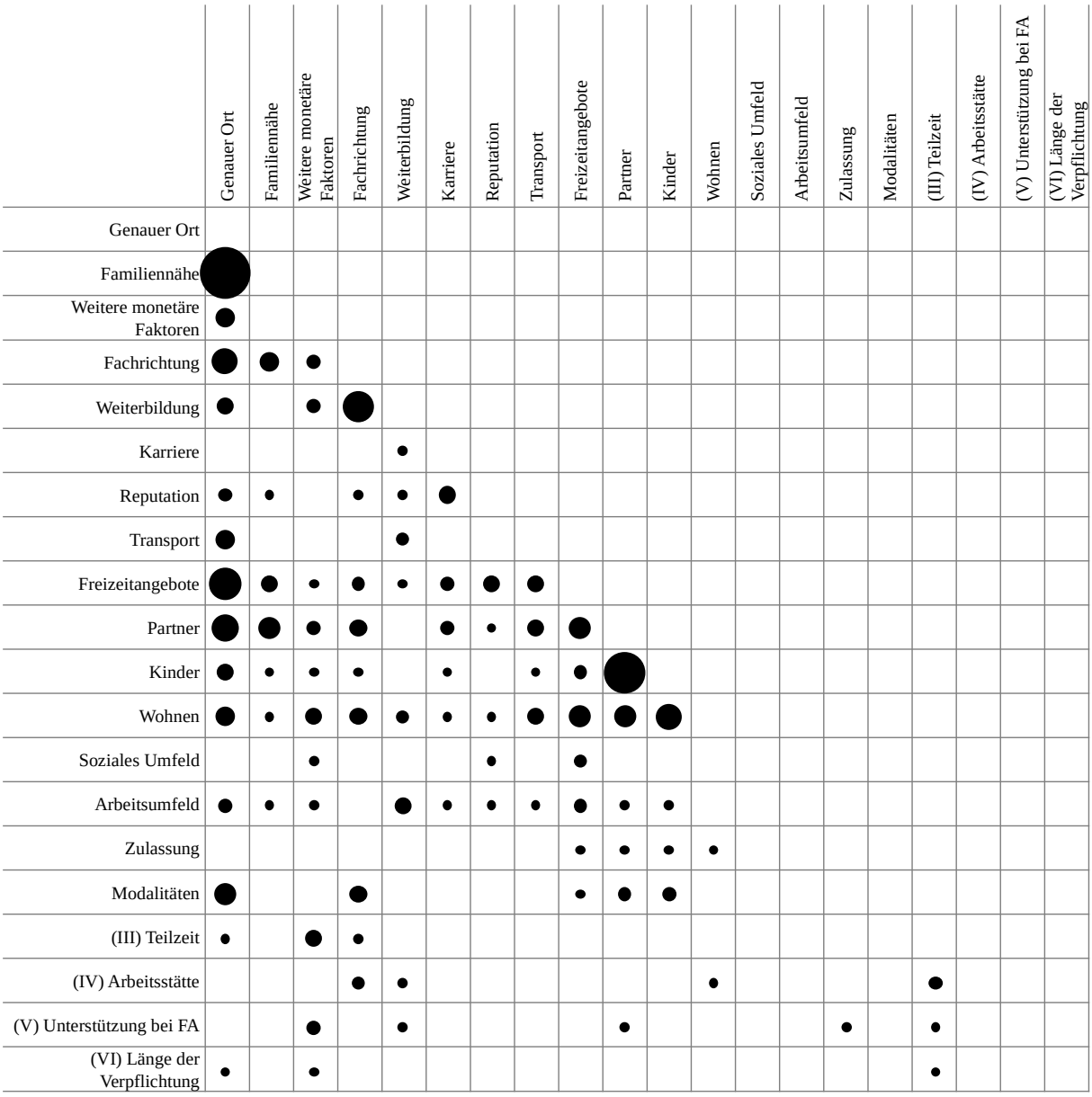


Figure 4: Gemeinsamnennungen zu qualitativen, wichtigen Faktoren für die Entscheidung der Verpflichtung (Mehrfachnennungen möglich)

Quelle: Eigene Erhebung

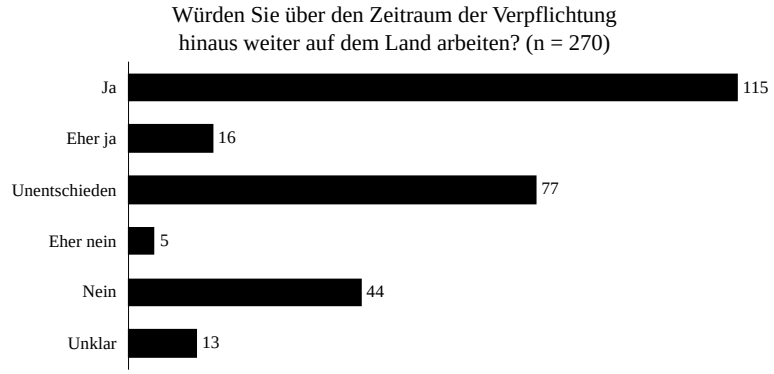


Figure 5: Eigeneinschätzung der Erhaltung als Landarzt

Quelle: Eigene Erhebung

Table 8: Begründungen für die Entscheidung auf dem Land zu bleiben (Mehrfachnennungen möglich)

Quelle: Eigene Erhebung

Kategorie	Ja	Eher ja	Unentschieden	Eher nein	Nein
Genauer Ort	11	1	6	0	4
Familiennähe	7	0	2	0	1
Weitere monetäre Faktoren	11	0	4	1	0
Fachrichtung	6	2	4	0	5
Weiterbildung	12	3	7	0	9
Karriere	4	1	2	0	2
Reputation	1	0	1	0	0
Transport	17	2	5	1	4
Freizeitangebote	14	2	7	1	12
Partner	20	1	18	0	2
Kinder	21	1	14	0	3
Wohnen	9	0	5	0	0
Soziales Umfeld	13	3	12	1	0
Arbeitsumfeld	28	6	15	1	0
Zulassung	0	1	0	0	0
Modalitäten	1	0	1	0	0
(III) Teilzeit	7	0	1	0	0
(IV) Arbeitsstätte	5	1	4	0	3
(V) Unterstützung bei FA	2	0	0	0	0
(VI) Länge der Verpflichtung	1	0	1	0	0
Total	**190**	**24**	**109**	**5**	**45**

Gegen die Erhaltung von Landärzten spricht am stärksten der Mangel an Freizeitangeboten (12 Nennungen) sowie an Weiterbildungsmöglichkeiten (9 Nennungen). Für Ersteres werden hauptsächlich Kulturangebote in der Stadt genannt, jedoch nicht präzisiert. Für die Weiterbildungsangebote berichten die Probanden häufig davon, dass sie bereits eine spezialisierte, klinische Wunschfachrichtung haben, für die es auf dem Land kaum Weiterbildungsangebote gibt, z.B. Neurochirurgie.

Die qualitative Analyse lässt sich insofern zusammenfassen, als dass die Heterogenität wie bei den vorhergehenden Parameterschätzungen und WTA-Berechnungen auch hier zu finden ist. Dort waren insbesondere weibliche, alleinstehende Medizinstudenten, die aus ländlichen Gebieten kommen und sich in früheren Studienabschnitten befinden, eher adressierbar, indem diese eine positive, aber realistische WTA und eine höhere Akzeptanzwahrscheinlichkeit hatten. Dieselben Gruppen geben weiterhin an, eher nach Ablauf der Verpflichtung weiterhin auf dem Land zu praktizieren. Ferner ist diesen für die Rekrutierung eher wichtig, den genauen Ort auszusuchen, der häufig in der Nähe bestehender Familie liegen soll.

5. Diskussion

Auf Basis der Ergebnisse des vorherigen Kapitels werden im Rahmen dieses Diskussionsteils einige Ergebnisse aus der Literatur bestätigt und ein realistischer Anwendungsfall für ein Verpflichtungsprogramm für Landärzte beschrieben.

Es wurde aufgezeigt, dass ein Verpflichtungsprogramm einen Teil der Medizinstudenten in Deutschland ansprechen kann. Besonders wichtig hierbei ist der genaue Ort der Anstellung im Rahmen der Verpflichtung; Faktoren wie Nähe zu Familie, Freunde und Partner sowie Freizeitangebote, die Erreichbarkeit der nächsten Großstadt und die weitere Infrastruktur auf dem Land sind relevant für die Entscheidung. Eine Vergütung ist wichtig im Rahmen der Verpflichtung, jedoch liegt diese im Bereich der durchschnittlichen Haushaltsausgaben oder niedriger. Generell sind Studenten, die sich verpflichten würden, tendenziell jünger, weiblich und alleinstehend als solche, die eine Verpflichtung ablehnen.

5.1. Reflexion des Literaturüberblicks

Der Literaturüberblick in Kapitel 2 ist in drei Teile gegliedert, wovon hauptsächlich der Erste nachfolgend mit den Ergebnissen der Umfrage verglichen werden soll. Die Thematiken des zweiten und dritten Teils sind ebenfalls grundlegend für diese Arbeit; jedoch kann aufgrund der Forschungsfrage und der Methodik dieser Arbeit keine direkte Reflexion beschrieben werden.

Viele der in der Literatur beschriebenen Faktoren, welche die Rekrutierung und Erhaltung von Landärzten beeinflussen, konnten sowohl in der Analyse der Parameterschätzungen und der WTAs als auch in der Auswertung der Freitextantworten aufgezeigt werden. Bei den individuellen Faktoren wurde beschrieben, dass sich jüngere Probanden wahrscheinlicher verpflichten (Kolstad, 2011). Im Rahmen dieser Arbeit hing das Alter weniger stark mit der Akzeptanzwahrscheinlichkeit zusammen, als der Studienabschnitt. Nichtsdestotrotz folgt diese Arbeit damit der Idee, dass mit höherem Alter, was z.T. mit einem späteren Studienabschnitt korreliert, sich Medizinstudenten eher binden. Eine ähnliche Relation zum Beziehungsstatus, wie sie an anderer Stelle beschrieben wurde (Roick et al., 2012), konnte nicht erkannt werden.

Ferner zeigten sich, ähnlich zu anderen Studien in Deutschland (Roick et al., 2012) Frauen interessierter an der Arbeit auf dem Land, während internationale Studien das Berufsbild des Landarztes eher männlich konnotieren.

Auch wurde in der qualitativen Analyse die Wichtigkeit des Partners und dessen Bedürfnisse sichtbar. Damit reiht sich diese Arbeit in eine Linie mit anderen Studien (z.B. Daniels et al., 2007; Han & Humphreys, 2006; Mayo & Mathews, 2006) ein, welche die Bedürfnisse, Arbeitsplatz und Integration des Partners als wichtigen Faktor für die Rekrutierung und Erhaltung von Landärzten beschreiben.

Außerdem zeigte sich wie bei anderen Studien eine ländliche Herkunft als relevanter Einfluss für die Präferenzen hin zur Arbeit als Landarzt. In dieser Arbeit wurde nur nach Herkunft gefragt, so dass dies an Ergebnisse bezüglich des Geburtsortes und des Ortes des Aufwachsens anknüpft (Geyman et al., 2000; Hancock et al., 2009; Laven & Wilkinson, 2003). Eine genauere Systematik, wie z.B. bei Hancock et al. (2009) in das Entwicklungsmodell, wurde nicht vorgenommen. Allerdings können Überlappungen von einigen Kategorien mit dem Entwicklungsmodell aufgezeigt werden. Viele Nennungen bei den Freitextfragen decken die Pfade der Familiarität, Gemeinschaft und zu einem gewissen Teil auch die der Selbstverwirklichung ab. Somit wird teilweise an Hancock et al. (2009) angeschlossen und das Auftreten der Pfade kann auch für den deutschen Kontext angedeutet werden.

Für die professionellen und organisatorischen Faktoren finden sich weitere Übereinstimmungen. Jedoch gilt zu beachten, dass diese Arbeit von bestehender Literatur abweicht, indem ein Verpflichtungsprogramm untersucht wurde. Bestehende Studien, darunter auch einige DCEs, haben häufig die reinen Präferenzen untersucht. Wenn eine finanzielle Komponente eingefügt wurde, so fand die Zahlung zeitgleich zur Arbeit auf dem Land statt. Somit wird erstmals der Fall einer sequentiellen Zahlung anstelle einer simultanen analysiert. Nichtsdestotrotz verhält sich die Zahlung im Rahmen des Verpflichtungsprogramms ähnlich wie die Zahlungen während der Arbeit auf dem Land in anderen Studien. Dort sind zwar die Zahlungen wichtig, jedoch nicht der einzige Faktor, den es zu berücksichtigen gilt (Kolstad, 2011; Scott et al., 2013). Aufgrund der Art und Weise, wie die WTA berechnet wird, kann nicht quantitativ belegt werden, dass Männern die Vergütung und deren Höhe wichtiger ist als Frauen (vgl. Szafran et al., 2001). Jedoch zeigt sich in der qualitativen Analyse, dass männliche Probanden im Vergleich zu den Probandinnen eher das Gehalt als weiteren wichtigen Faktor sehen.

Darüber hinaus wurde in der bestehenden Literatur gezeigt, dass Arbeitszeit (Daniels et al., 2007; Hegney et al., 2002; Kasch et al., 2016) und Bereitschaftsdienste (Scott et al., 2013; Ubach et al., 2003) einen relevanten Einfluss auf die Erhaltung von Landärzten haben. Dies konnte im Rahmen dieser Arbeit nicht gezeigt werden. Nachdem Bereitschaftsdienste bereits in der Vorstudie eliminiert wurden, wurden diese nur ein einziges Mal im qualitativen Teil der Hauptstudie erwähnt. Auch wurde die Arbeitszeit nur im Rahmen des allgemeinen Arbeitsumfeld zusammen mit anderen Faktoren, wie die Beziehungen zu Kollegen und Vorgesetzten oder die Ausstattung der Arbeitsstätte, angesprochen. Fraglich ist, woher diese Abweichung kommt. So basiert diese Arbeit auf Medizinstudenten, die abgesehen von den Pflichtpraktika während des Studiums keinen Berührungspunkt mit den Themen Arbeitszeit oder Bereitschaftsdienst haben. Weiterhin konzentriert sich diese Arbeit ausschließlich auf Deutschland, während andere Studien vorwiegend aus angelsächsischen Ländern kommen. Allgemein zeigt sich hier eine Lücke in der Literatur, bei der das Arbeitsumfeld für Landärzte nur wenig untersucht wurde. Entweder findet die Analyse auf einem sehr detaillierten Level statt, bei dem einzelne Faktoren wie die Arbeitszeit auf ihren Einfluss auf Rekrutierung und Erhal-

tung untersucht werden (z.B. Daniels et al., 2007; Hegney et al., 2002; Kasch et al., 2016), oder auf einem generalisierten, sehr hohem Level, indem eine latente Variable in Form von Arbeitszufriedenheit o.Ä. (z.B. Cooper et al., 1989; Grembowski et al., 2005) herangezogen wird. Die genaue Beziehung der Einzelfaktoren wie der Arbeitszeit zur Arbeitszufriedenheit und schlussendlich zur Rekrutierung und Erhaltung von Landärzten ist wenig untersucht. Insbesondere die Rolle dieser Faktoren sowie die möglichen Moderator- sowie Mediator-Variablen stellen ein weiteres Feld für die Forschung dar. Nichtsdestotrotz konnte in der qualitativen Analyse gezeigt werden, dass solche Faktoren im Speziellen für die Erhaltung von Landärzten wichtig sind.

Außerdem konnte die Relevanz der Weiterbildung für Landärzte bestätigt werden. Durch die Eigenschaft (V) Unterstützung bei der Facharztweiterbildung wurde dieser Aspekt mit in die Betrachtung aufgenommen. Es zeigt sich deutlich der positive Effekt einer Unterstützung bei der Weiterbildung auf die Rekrutierung bzw. Erhaltung von Landärzten. Überdies stellt die Weiterbildung, insbesondere die zum Facharzt, auch im qualitativen Teil z.T. einen Grund dagegen, Landarzt zu werden, dar. Wie in anderen Studien (Hegney et al., 2002; Kolstad, 2011) beschrieben, gilt es dies zu adressieren, um dem Landarztmangel entgegenzuwirken.

In der Literaturübersicht werden innerhalb der Systematik noch systemische und institutionelle Faktoren beschrieben. Da diese Arbeit sich ausschließlich auf Deutschland konzentriert, können solche Faktoren nur peripher behandelt werden. In diesem Bereich ist jedoch das wichtigste Ergebnis, dass im Gegensatz zu amerikanischen Studien (Daniels et al., 2007; Rosenblatt et al., 1992) keine Universität oder medizinische Fakultät ausfindig gemacht werden konnte, die besonders „landarztfreundlich“ ist. Um weitergehend andere Faktoren, wie Bürokratie, Dezentralisierung der Organisation oder das Krankenversicherungssystem und deren Einfluss zu untersuchen, müsste eine Studie entweder länderübergreifend durchgeführt werden oder es müsste ein anderes Experiment dafür aufgesetzt werden.

Auf Ebene der soziokulturellen Faktoren können hauptsächlich die Ergebnisse anderer Studien bezüglich der Wichtigkeit der Integration in die ländliche Gemeinschaft bestätigt werden. Insbesondere bei der qualitativen Erhebung zur Erhaltung, aber auch für die Rekrutierung wurde die soziale Integration in die ländliche Gemeinschaft als wichtiger Faktor genannt. Dieser trat unabhängig von demografischen oder sonstigen erhobenen Merkmalen auf und bestätigt damit Ergebnisse anderer Studien (Hancock et al., 2009). Gleiches gilt für die Integration des Partners, welche bereits zuvor in der wissenschaftlichen Literatur untersucht wurde (Han & Humphreys, 2006).

Ferner wird auch die Rolle des Faktors Prestige kurz umrissen. Im qualitativen Teil traten Freitextantworten in der Kategorie Reputation und Anerkennung insbesondere bei männlichen Probanden auf und passen somit zu bereits bestehenden Studien (Dussault & Franceschini, 2006).

Obendrein zeigt diese Arbeit auch die Relevanz von Freizeitangeboten auf, besonders für die Entscheidung, nicht auf das Land zu gehen bzw. dort zu bleiben. Andere Studien (Dussault & Franceschini, 2006) haben bereits zuvor diesen Effekt diskutiert. Im Rahmen dieser Arbeit stachen jedoch besonders „Kulturangebote“ als Freizeitangebote heraus, die für bestimmte Probanden ländliche Gebiete weniger attraktiv machten. Andere Studien (Dionne et al., 1987) hatten stattdessen Restaurants als Freizeitangebote betrachtet. Zudem stellt die zugrundeliegende Arbeit auch heraus, dass es auf dem Land nicht per se an Freizeitangeboten mangelt, sondern diese in anderer Form auftreten können. Einige Probanden erwähnten die Natur als positives Freizeitangebot, die sie zum Bleiben bewegen würde. Damit löst sich diese Arbeit von der in der Literatur vorherrschenden Konnotation von Freizeitangeboten als rein kulturelle oder soziale Angebote. Ferner impliziert dies, dass für bestimmte Personengruppen ländliche Gebiete einen höheren Freizeitwert als städtische Gebiete aufweisen können. Weiterführende Studien könnten an dieser Stelle verschiedene Freizeitangebote in der Natur (z.B. Berge, Seen) in eine Systematik einordnen und einen Freizeitwert bestimmen.

Bezüglich der Gegenüberstellung von regulatorischen und administrativen Maßnahmen sowie finanziellen und professionellen Anreizen, kann auch kein Beweis erbracht werden, dass finanzielle und professionelle Anreize überlegen wären. Jedoch konnte im Rahmen des DCEs gezeigt werden, dass auch für die Situation in Deutschland ein realistischer Preisrahmen für junge Ärzte bzw. hier Medizinstudenten abgebildet werden kann (Yong et al., 2018). Damit stellt diese Arbeit einen weiteren Schritt in die Richtung dar, den Landärztemangel nicht nur bezogen auf bereits praktizierende Ärzte wissenschaftlich zu untersuchen.

5.2. Handlungsempfehlungen

Das folgende Unterkapitel gibt Handlungsempfehlungen für die konkrete Behebung des Landärztemangels durch ein Verpflichtungsprogramm. Diese sind aufgrund der Struktur des deutschen Gesundheitssystem an den deutschen Staat gerichtet.

Um Handlungsempfehlungen zur Behebung eines Landärztemangels zu geben, gilt es diesen zunächst von einer ökonomischen Perspektive zu verstehen. Entgegen der öffentlichen Diskussion über einen Ärztemangel wurde im Rahmen dieser Arbeit nicht seine Existenz untersucht. Wie bereits vorhergehend diskutiert, leidet Deutschland nicht an einem generellen Ärztemangel. Stattdessen mangelt es insbesondere an der medizinischen Grundversorgung auf dem Land, daher handelt es sich um einen Landärztemangel. An vorhergehender Stelle wurde dessen Existenz bereits belegt. Grundsätzlich sollte vom Standpunkt der ökonomischen Theorie kein regionaler Ärztemangel existieren, solange ein effizienter Markt für die Gesundheitsversorgung besteht. Das Ungleichgewicht zwischen Angebot und Nachfrage, d.h. ein Unterangebot auf dem Land, würde die Preise für die Leistungen der Landärzte und damit deren Löhne steigen lassen, wodurch mehr Ärzte aufs Land wechseln würden. Durch das höhere Angebot sinken der Preis der Gesundheitsleistungen und die Löhne wieder bis ein Gleichgewicht

zwischen Angebot und Nachfrage erreicht wäre. Allerdings ist der Markt für Gesundheitsversorgung nicht effizient im Sinne der ökonomischen Theorie, da er vom deutschen Staat stark reguliert wird. So sind die Preise für Behandlungen über den EBM und die GOÄ und ein Großteil der Löhne über Tarifverträge festgelegt. Außerdem bestehen durch die Studienzulassungsbeschränkungen, den stark regulierten Approbationsprozess und den Prozess der Kassenzulassungen Markteintrittsbarrieren. Weiterhin wirken sich auch eine Vielzahl von nicht-monetären Faktoren auf die Ortswahl von Ärzten aus.

Diese Ineffizienz im Markt kann in der aktuellen Situation nicht durch den freien Markt gelöst werden. Aufgrund des fehlenden Preismechanismus und der Markteintrittsbarrieren wird für die limitierte Anzahl an Ärzten in Deutschland kein zusätzlicher Anreiz geschaffen, auf dem Land zu praktizieren. Bei einer Deregulierung des Marktes durch die Aufhebung der Preisbindung und Markteintrittsbarrieren wäre es jedoch fraglich, ob dies die Lage auf dem Land stark verbessern würde. Da die Zahlungskraft der Patienten auf dem Land durchschnittlich geringer ist und städtische Regionen auch in anderen Aspekten für Ärzte attraktiver sind (Sundmacher & Ozegowski, 2016), könnte selbst vermutlich eine Deregulierung den Landärztemangel nicht vollständig lösen. Weiterhin wäre dies eine Reform, die das deutsche Gesundheitssystem im Kern verändern würde. Es ist zutiefst unwahrscheinlich, dass diese vom gesellschaftlichen Konsens getragen und mit dem Grundgesetz vereinbar wäre (Wissenschaftlicher Dienst des Deutschen Bundestages, 2015). Die weitere Beurteilung einer solchen Maßnahme entzieht sich daher dem Rahmen dieser Arbeit.

Stattdessen kann dieses Ungleichgewicht mit einer verstärkten und besseren Regulierung des Marktes zumindest verringert werden.

Erstens hat der deutsche Staat dafür die Möglichkeiten im Rahmen der Gesetzgebung und Gestaltung des Gesundheits- und Ausbildungssystems. In seiner Charakteristik als Ungleichverteilung von Ärzten bildet das Thema des Landärztemangels einen Berührungspunkt der Gesundheitspolitik mit der Bildungspolitik. Dabei beeinflusst erstere primär die Verteilung der Ärzte. Dennoch kann auch die Bildungspolitik dazu beitragen, das Berufsbild eines Landarztes attraktiver zu gestalten. Die Kompetenzen dafür liegen teilweise auf der Ebene des Bundes, wie z.B. die Approbationsordnung, und teilweise auf Landesebene, wie z.B. spezifische Hochschulzulassung und Kassenzulassung.

Zweitens zieht der Staat auch einen großen Nutzen aus der Behebung des Landärztemangels. Primär wirkt sich ein Landärztemangel auf die medizinische Grundversorgung und damit direkt auf die Gesundheit der auf dem Land lebenden Bürger aus. Folglich wird durch die Behebung des Landärztemangels die Gesundheit der Bürger verbessert, deren allgemeines Wohl erhöht (Lamu & Olsen, 2016) und die Ungleichheit zwischen Stadt und Land innerhalb der deutschen Gesellschaft reduziert. Obendrein lässt sich auch aus dem ethisch-moralischen Diskurs eine Verpflichtung des Staates bzw. der handelnden Politiker ableiten, das Wohl der Bürger durch die Behebung des Landärztemangels zu steigern. Die genaue Argumentationslinien hierfür können an anderer Stelle gefunden werden (z.B. Childress et al., 2002) und befinden sich außerhalb des Rahmens dieser Arbeit.

Ferner lässt sich auch die Pflicht, dem Bürger ein funktionierendes Gesundheitssystem zur Verfügung zu stellen, im deutschen Grundgesetz, genauer gesagt in Art. 2 Abs. 2 Satz 1 GG in Verbindung mit Art. 20 Abs. 1 GG, finden (Wissenschaftlicher Dienst des Deutschen Bundestages, 2015). Daneben verursacht der Landärztemangel eine Reihe an Folgekosten, die zum Großteil vom Staat getragen werden. Wie im Literaturüberblick diskutiert, reduziert eine gute medizinische Grundversorgung das Auftreten und die Schwere von vielen Krankheiten. Außerdem wird die Anzahl der vermeidbaren Krankenhausaufenthalte verringert. Dadurch entfallen hohe Folgekosten und das Gesundheitssystem muss weniger Kapazität vorhalten. Der Staat trägt indirekt einen Teil dieser Kosten, z.B. durch den Bundeszuschuss zur GKV oder die Investitionskosten für zusätzliche Kapazitäten im Gesundheitssystem. Somit hat der deutsche Staat Mittel und Motiv zur Behebung des Landärztemangels.

Anwendung eines Verpflichtungsprogramms: Das „Umgekehrte Rentensystem“ basierend auf der vorhergehenden Diskussion und den Ergebnissen dieser Arbeit soll im Folgenden ein beispielhaftes Programm entworfen werden, welches es schafft, den Landärztemangel in Deutschland zu beheben. Auf Basis der Ergebnisse des DCEs ist insbesondere die Flexibilität des Programms wichtig. Dies gilt sowohl für die Dauer, den Ort als auch die genauen Modalitäten. Diese Punkte werden im nachfolgend beschriebenen Programm aufgegriffen. Dabei werden insbesondere die Eckpunkte Initiator, Adressaten und das Programm selber angesprochen. Das Programm orientiert sich am deutschen Rentensystem, wobei es jedoch den Zeitpunkt von Zahlung und Arbeit umdreht. Der Student kann während des Studiums monatliche Zahlungen abrufen, für die dann Punkte aufgebaut werden. Diese Punkte können dann nach der Approbation durch das Praktizieren auf dem Land bundesweit wieder abgebaut werden, wobei der genaue Abbaufaktor von der Art der Tätigkeit anhängt. Wie im vorherigen Kapitel beschrieben, wäre somit die Möglichkeit Teilzeit zu arbeiten gut integrierbar. Akzeptable Möglichkeiten, die Verpflichtung zu kündigen, müssen ebenfalls angeboten werden. Aufgrund seiner Struktur wird dieses Programm auch nachfolgend als „Umgekehrtes Rentensystem“ bezeichnet. Der Vorschlag zeichnet sich durch ein hohes Maß an Flexibilität und einer genauen Steuerbarkeit aus. Weiterhin führt dieser Entwurf durch die Entkopplung von Zahlung- und Verpflichtungszeitraum bei einer Einstellung des Programms nicht dazu, dass die Versorgung zeitnah stark zurückgeht, wie es bei anderen finanziellen Anreizprogrammen der Fall ist (Bärnighausen & Bloom, 2009). Zusätzlich kann die Attraktivität des Programms noch mit weiteren Angeboten, wie z.B. der Facharztweiterbildungsunterstützung, erhöht werden. Exemplarisch wurde der Effekt dieser Attraktivitätssteigerungen bereits in Kapitel 4 demonstriert.

Im Vergleich zum Status Quo ohne das Verpflichtungsprogramm, werden im Rahmen des „Umgekehrten Rentensystems“ durch eine größere Anzahl an Landärzten die Gesundheitsversorgung des Bürgers auf dem Land verbessert und dadurch Folgekosten eingespart. Abbildung 6 zeigt einen simplifizierten Vergleich der Situation mit und ohne Verpflichtungsprogramm in Deutschland. Der Investitionscharakter des Programms wird hierbei deutlich sichtbar.

Initiator

Ein wichtiger Eckpunkt des beschriebenen Programms ist die Frage nach dem Initiator. Auf der öffentlichen Seite können Bundes- oder Landesebene das Programm initiieren. Aufgrund der obigen Diskussion wird das Szenario für eine privatwirtschaftliche Unternehmung nicht analysiert. Die wichtigsten Entscheidungspunkte hierbei sind der geografische Rahmen, die Kompetenzen für die gesetzlichen Rahmenbedingungen, politischer Einfluss, die inhaltliche Kompetenz sowie die benötigten Mittel.

Der geografische Rahmen für das Programm sollte die gesamte Bundesrepublik sein. Der Landärztemangel ist kein Problem, welches auf ein Bundesland beschränkt ist, sondern betrifft alle 13 Länder, die keine Stadtstaaten sind (KBV, 2016). Da dieses Problem somit fast die gesamte Bundesrepublik betrifft, liegt zumindest eine Koordination auf Bundesebene nahe. So zeigt sich im Berufsmonitoring der KBV (2018a), dass mit Ausnahme von Bremen und dem Saarland das Herkunftsbundesland als bevorzugter Arbeitsort wahrgenommen wird. Generell wird ein Arbeitsort in der Nähe des Herkunftsortes oder des Studienortes vorgezogen (KBV, 2018a). Wie bereits in Kapitel 4 angesprochen, studieren eine Vielzahl von Studenten nicht in dem Bundesland, in welchem sie aufgewachsen sind. Ein Programm auf Landesebene hingegen, wäre damit nur attraktiv für Studenten, die in diesem Bundesland studieren und dort auch bleiben wollen. Da medizinische Fakultäten und die Zahl der Medizinstudienplätze ungleich über die Bundesländer verteilt sind, ist es in vielen Fällen nicht möglich oder vom Studenten gewünscht, in einem Bundesland zu bleiben. Zusätzlich lässt sich so die Einheit innerhalb Deutschlands demonstrieren und eine Konkurrenzsituation zwischen den Gliedstaaten, wie in den USA, vermeiden (de Vise, 1973).

Bezüglich der Kompetenzen fällt das Programm in den Bereich von Bund und Ländern, wodurch eine Kooperation beider nötig wäre. So liegt für die Ausbildungsförderung, die Gesundheitspolitik sowie die Hochschulzulassung und -abschlüsse der Fall der konkurrierenden Gesetzgebung vor. Hierbei erlässt der Bund Gesetze, die von den Ländern mit Abwandlungen in eigenes Recht überführt werden. Im Falle der Ausbildungsförderungen und Gesundheitspolitik findet ein hohes Maß an Koordination statt. Bei ersterem übernimmt der Bund z.B. die Zahlung der BAföG Förderungen. Bei letzterem folgen die Länder größtenteils der Bundesgesetzgebung und dem Bundesministerium für Gesundheit sowie der Koordination durch die Gesundheitsministerkonferenz der Länder, sodass die Situation deutschlandweit ähnlich ist (Bundesministerium für Gesundheit, 2020). Weiterhin räumt das Grundgesetz den Ländern die sogenannte Kulturhoheit ein (siehe Art. 74 Abs. 1 Nr. 13; 91 a Abs. 1 Nr. 1 GG sowie Art. 30 GG). Im Rahmen dieser stellt die Hochschulpolitik eine wichtige Kompetenz der Länder dar. Zwar hat der Bund die Möglichkeit, die Hochschulzulassung und -abschlüsse zu regeln, allerdings machen die Länder hier auch von ihren Kompetenzen Gebrauch und haben eigene Ländergesetze, wie sich auch bei der Situation der Landarztquote zeigt. Aufgrund des deutschlandweiten Rahmens des Programms wäre es jedoch sinnvoll, dass Bund und Länder hierbei kooperieren.

Für die inhaltliche Ausgestaltung des Programms stehen Bund und Ländern auf beiden Ebenen zahlreiche Kooperationspartner zur Verfügung. Parallel zur föderalen Struktur des deutschen Staates, finden sich ähnliche Strukturen bei der KBV und den KVen sowie der Bundesärztekammer (BÄK) und den Landesärztekammern (LÄK). Somit könnte das Programm unter Federführung des Bundes aufgesetzt werden, welcher auf dieser Ebene Unterstützung durch die KBV und die BÄK erhält. Bei der operativen Ausführung können die Länder auf ihre bestehenden Systeme, die u.a. auch für die Förderung durch BAföG verwendet werden, zurückgreifen. Für inhaltliche Anliegen stehen hier die LÄKs zur Verfügung und die Bedarfsplanung für den Ort der Verpflichtung kann durch die KVen übernommen werden. Wichtig hierbei ist jedoch wieder die Aggregation der Information auf Bundesebene, um einen Überblick zu gewährleisten. Genauere Details, wie z.B. welche Ministerien – etwa das Bundesministerium für Bildung und Forschung oder das Bundesministerium für Gesundheit –involviert werden, können in späteren Planungsschritten konkretisiert werden.

Die Mittel für das Verpflichtungsprogramm sind aus dem Bundeshaushalt realisierbar, wobei eine Beteiligung der Länder denkbar wäre. Da die genaue Planung den Rahmen dieser Arbeit übersteigt, soll lediglich die Größenordnung in einer Maximalrechnung veranschaulicht werden. Bei aktuell 96000 Medizinstudenten liegen die rechnerischen Höchstkosten bei 2,3 Milliarden Euro jährlich plus evtl. Verwaltungskosten, wenn jeder dieser Medizinstudenten in Deutschland im Monat eine Vergütung von 2000 € erhalten würde. Natürlich wären die realen Kosten deutlich niedriger, da in keinem realistischen Szenario alle Medizinstudenten teilnehmen würden bzw. benötigt werden würden, um einem Landärztemangel entgegen zu wirken. Auch ist eine durchschnittliche Vergütung von 2000 € im Monat unrealistisch hoch. Die Verwaltungskosten können ähnlich zu denen der BAföG Förderungen angenommen werden und sind bei dieser Höhe vernachlässigbar gering (Deutscher Bundestag, 2017). Nichtsdestotrotz liegt diese Summe noch unter den jährlichen Ausgaben von 2,7 Milliarden Euro für die BAföG Förderungen (Stand 2018) (Statistisches Bundesamt, 2019) und den 14,5 Milliarden Euro an GKV Zuschüssen (Stand 2020), die der Bund leistet (Bundesministerium der Finanzen, 2020). Bei der Größe des aktuellen Bundeshaushalts 2020 von 484,5 Milliarden Euro das mögliche Verpflichtungsprogramm weniger als 0,5% der Ausgaben dar (Bundesministerium der Finanzen, 2020) Außerdem würde

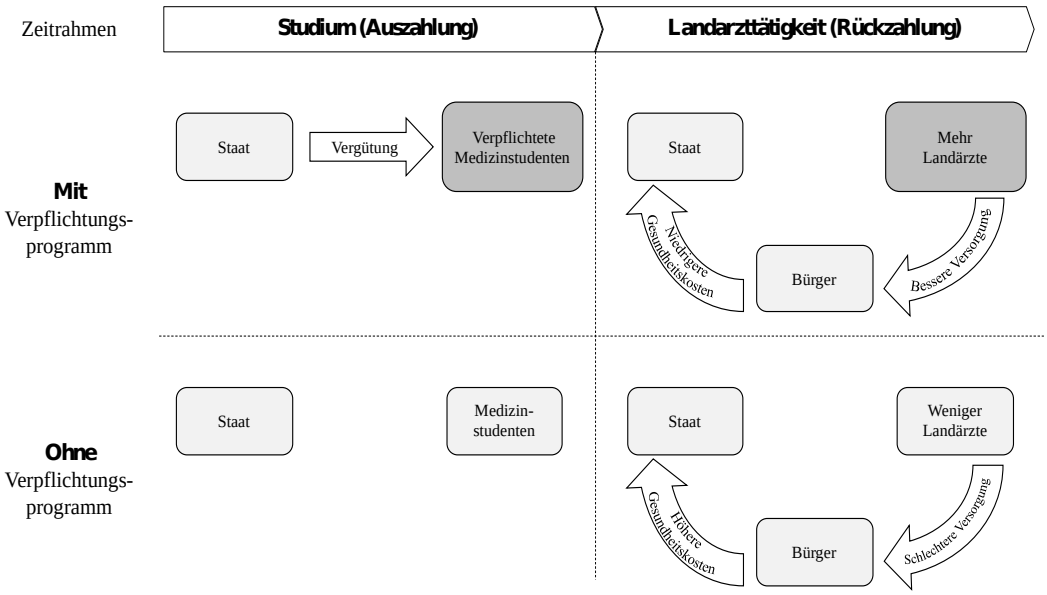


Figure 6: Ziel des "Umgekehrten Rentensystems"

Quelle: Eigene Abbildung

der Bund auch den größten Nutzen dieses Programms tragen. Den Ausgaben in Form der monatlichen Vergütungen steht ein Nutzen in Form von reduzierten Gesundheitskosten gegenüber, deren Höhe an späterer Stelle diskutiert wird. Auch würden Länder und Kommunen Einsparungen durch weniger benötigte Investitionen in Krankenhäuser realisieren. Somit stellt ein Landärzteverpflichtungsprogramm eine Investition für Bund und Länder dar, die hilft, an späterer Stelle Geld durch verringerte Folgekosten zu sparen.

Als Initiator und Träger eines Verpflichtungsprogramms sollten Bund und Länder kooperieren, um deutschlandweit ein tragbares Programm aufzusetzen.

Adressaten

Die Adressaten des Programms sind zunächst alle Medizinstudenten an deutschen medizinischen Fakultäten, wobei einige Ausnahmen getroffen werden müssen. Weiterhin gilt es sicherzustellen, dass die richtige Zahl an Studenten rekrutiert wird. In Anlehnung an Fruen et al. (1980) reicht es, genug Studenten zu verpflichten, um den Landärztemangel zu beheben. Eine Verpflichtung weiterer Studenten ist zu diesem Zweck nicht zielführend. Schätzungen der genauen Anzahl an benötigten Landärzten sind rar und geben keine eindeutigen Zahlen an.

Für die Teilnahme am Verpflichtungsprogramm kommen generell alle Medizinstudenten in Frage, welche in Deutschland geplant ihre Approbation erhalten werden und somit als Landärzte tätig werden können. Aus praktischen Gesichtspunkten sollten die Fälle von ausländischen Staatsbürgern separat untersucht werden, um die Tauglichkeit sicherzustellen. Diese besteht einerseits aus ausreichenden Deutschkenntnissen. Darüber hinaus gilt es sicherzustellen, dass die Abarbeitung der Punkte garantiert werden kann. Somit sollten nur ausländische Studenten gefördert werden, die sich nach Abschluss des Studiums voraussichtlich in Deutschland aufhalten dürfen. Zusätzlich sind Einschränkungen bezüglich der Sinnhaftigkeit der Teilnahme notwendig. Dass ein Student im letzten Studienabschnitt teilnimmt, scheint wenig sinnvoll, da lediglich eine kurze Verpflichtungsdauer angemessen wäre. Außerdem kann noch ein separater Fall für deutsche Staatsbürger, die Medizin im Ausland studieren, untersucht werden. Diese würden auch in das Profil passen, jedoch obliegt es dem deutschen Staat zu entscheiden, ob die Ausbildung im Ausland eine akzeptable Qualität hat und, ob Geldströme ins Ausland wünschenswert sind.

Um die Effizienz des Verpflichtungsprogramms sicherzustellen, sollten besonders alleinstehende, weibliche Medizinstudenten, welche vom Land kommen und sich noch am Anfang des Studiums befinden, angesprochen werden. Die Ergebnisse dieser Arbeit haben gezeigt, dass diese vier Merkmale einen hohen Nutzwert für die Teilnahme an einem Landärzteverpflichtungsprogramm aufweisen. Ferner wurde gezeigt, dass die Gruppe der Medizinstudenten sehr heterogen ist und dadurch für die effizientere Ansprache eine Segmentierung stattfinden muss. So wird es nicht möglich sein, allen Medizinstudenten die passenden Anreize zu setzen, um sich zu verpflichten. Einige von ihnen, insbesondere Männer, die aus der Großstadt kommen und sich am Ende ihres Studiums befinden, werden sich wahrscheinlich in keinem Programmdesign, welches im Rahmen dieser Arbeit erdacht würde, verpflichten. So sollten unter Gesichtspunkten der Fairness und Gleichberechtigung die Teilnahme allen Medizinstudenten, die die oben beschriebenen Merkmale besitzen, offenstehen. Nichtsdestotrotz kann die gezielte Akquise von Teilnehmern durch die Segmentierung, die ein

Ergebnis dieser Arbeit darstellt, effizienter gestaltet werden.

Programm

In der genauen Gestaltung des Programms zeigen sich seine Vorteile durch eine genaue Bedarfssteuerung und die nötige Flexibilität, die von den Medizinstudenten gewünscht wird.

Wie eingangs beschrieben, können Teilnehmer des Programms während des Studiums eine monatliche Vergütung abrufen, für die im Gegenzug Punkte aufgebaut werden. Diese Punkte müssen dann nach der Approbation in einem angemessenen Zeitrahmen durch die Arbeit als Landarzt wieder abgebaut werden. Die genauen Umrechnungsfaktoren der Vergütung zu Punkten und der Punkte zu Arbeitsleistungen müssen weitergehend noch bestimmt werden, jedoch wird angenommen, dass die erbrachte Arbeitsleistung einen positiven Effekt für den deutschen Staat und seine Bürger bringt. Bei einer Studienzeit von sechs Jahren und drei Monaten (75 Monate) für ein Medizinstudium in Deutschland und einer Monatshöchstfördersumme von 2000 € , würde sich somit eine maximale Vergütung von 150000 € pro Medizinstudenten ergeben. Die gegenüberstehenden Gesundheitskosten lassen sich deutlich schwieriger berechnen, jedoch soll stattdessen die Größenordnung aufgezeigt werden. Das Statistische Bundesamt veranschlagt 4712 € an Gesundheitskosten pro Bundesbürger und Jahr (Statistisches Bundesamt, 2019). Basierend auf Zahlen der BÄK kommen auf jeden berufstätigen Arzt 207 Einwohner, d.h. potentielle Patienten (Bundesärztekammer, 2019). Damit stehen dem durchschnittlichen Arzt jährlich über 975000 € an Gesundheitsausgaben gegenüber. Bei einer Verpflichtung von drei Jahren, wie sie im Rahmen des DCEs untersucht wurde, oder noch länger, sollte es möglich sein, die Vergütung in der Höhe von 150000 € in Einsparungen von Folgekosten zu amortisieren. Dazu kommt, dass der Altersdurchschnitt auf dem Land höher ist, so dass hier mit deutlich höheren Pro-Kopf-Gesundheitskosten gerechnet werden kann. Ein weiterer Indikator, welcher in Unterkapitel 2.2 beschrieben wurde, ist die Rechnung für Brasilien, wo pro Kopf und Jahr über 200000 US Dollar gespart werden konnten. Somit kann trotz der Abwesenheit genauer Zahlen vermutet werden, dass sich das Verpflichtungsprogramm allein auf finanzieller Ebene rechnet. Weitere Implikationen auf nichtmonetärer Ebene, wie ethisch-moralische Handlungsanweisungen und andere politische Überlegungen sind in dieser Rechnung nicht berücksichtigt.

Als wichtiger Faktor für die Teilnahme wurde im Rahmen der Umfrage der genaue Ort genannt. Die KBV und die KVen berechnen und steuern die Versorgung in sogenannten Planungsgebieten alle zwei Jahre, wobei auch begrenzte Prognosen für die Zukunft getroffen werden (KBV, 2020). Somit kann je nach Versorgungsgrad der Region den Studenten schon vor Abschluss des Studiums ein Ort angeboten werden. Die Genauigkeit hängt hierbei jedoch von der medizinischen Fachrichtung ab. Für Allgemeinmediziner ist die Planung am detailliertesten und kennt 883 Planungsbereiche in Deutschland (KBV, 2020). Somit sollte, selbst wenn der Wunschort nicht verfügbar ist, ein Planungsbereich in annehmbarer Nähe gefunden werden können. Für andere Fachrichtungen, die auch als Landärzte dienen können, wie z.B. Kinder- und Jugendärzte, findet die Planung wenig granular in 361 Planungsbereichen statt (KBV, 2020). Jedoch sollte bei ausreichender Güte der Prognose auch hier eine Planung über den Zweijahreshorizont möglich sein. Außerdem kennt das Planungssystem eine gewisse Flexibilität, die für die bessere Erfüllung des Wunschortes genutzt werden kann. Über den Wunschort kann eine breite Anzahl von weiteren Faktoren, wie Freizeitangebote und Familiennähe, abgedeckt werden. Bei der Auswahl der Teilnehmer des Verpflichtungsprogramms sollte bereits nach dem Wunschort gefragt werden, so dass zu einem späteren Zeitpunkt Planungskonflikte vermieden werden können.

Ferner kann das Programm auch andere Anreize setzen. So können weitere Angebote wie die Unterstützung bei der Facharztweiterbildung oder Auslandsaufenthalte mit angeboten werden. Die Entscheidung, ob dies in Punkten verrechnet wird oder nicht, kann fallabhängig getroffen werden.

Ein weiterer wichtiger Faktor sind gerechte Vertragsmodalitäten, wie Kündigungskonditionen und Aufschübe. Generell kann das Programm so gestaltet werden, dass für den Fall, in dem das Studium nicht in der Regelstudienzeit beendet wird, die Zahlung der Vergütung fortgesetzt wird und dafür weiterhin Punkte aufgebaut werden. Auch für den Fall, dass die Punkte nicht direkt nach dem Ende des Studiums abgearbeitet werden oder in der Abarbeitungszeit Pausen benötigt sind, kann hier in Maßen Flexibilität dem verpflichteten Landarzt gegenüber gezeigt werden. Unter besonderen Umständen, wie z.B. Elternzeit, können auch andere Lösungen gefunden werden. Für den Kündigungsfall sollte die erhaltene Vergütung zurückgezahlt werden. Wenn nötig, kann diese gestückelt werden. Auch besteht die Möglichkeit einen Strafzins aufzuerlegen, um finanzielle Arbitragehandlungen unattraktiv zu machen. So geht selbst im Falle der Kündigung der Verpflichtung dem Staat lediglich Liquidität über die Dauer der Verpflichtung verloren. Zusätzlich können noch Vorkehrungen für Spezialfälle getroffen werden. Für den Fall, dass ein verpflichteter Student derartig talentiert ist, dass er in der medizinischen Forschung der Gesellschaft einen größeren Mehrwert bringen würde, können Universitäten und andere Forschungseinrichtungen „Freifahrtscheine“ bekommen. Damit könnte eine bestimmte Anzahl an Studenten ohne Strafzahlung aus der Verpflichtung befreit werden, damit diese sich in der Forschung engagieren. Für die medizinische Forschung außerhalb von Universitäten und anderen staatlichen Einrichtungen, im Speziellen in der Pharmaindustrie, sollte das Investment, einen talentierten Studenten aus einer Verpflichtung „rauszukaufen“, ein akzeptables Investitionsszenario darstellen. Somit sollten mögliche Innovationsverluste im Bereich der Medizin durch die Verpflichtung minimiert werden. Nach der Initiierung des Programms kann dann anhand von historischen Zahlen die Quote an Studenten, die die Verpflichtung auflösen, bestimmt und in die Planung miteinbezogen werden. Zusammenfassend für diese Beschreibung

gibt Abbildung 7 einen Überblick.

Weitere Überlegungen

Zur Bekämpfung des Landärztemangels in Deutschland stellt das vorhergehend beschriebene „Umgekehrte Rentensystem“ ein wichtiges Element dar, welches jedoch in eine nationale Strategie zur Verbesserung der Gesundheitsversorgung auf dem Land eingebettet werden sollte. Durch die Schaffung von Anreizen, um die Arbeit als Landarzt attraktiver zu machen, wird zwar ein Teil des Problems behoben, jedoch werden noch weitere Maßnahmen benötigt.

Neben dem Verpflichtungsprogramm selber sollten in diesem Rahmen auch die Arbeitsbedingungen von Landärzten verbessert werden. Das Ziel hierbei sollte sein, die Erhaltung von Landärzten zu erhöhen, um die Notwendigkeit von ständigen Rekrutierungsmaßnahmen zu verringern. Ein Teil der möglichen Maßnahmen wurde im Rahmen des Literaturüberblicks dieser Arbeit angesprochen und beinhaltet Möglichkeiten, wie eine Verminderung des administrativen Aufwandes, eine gute Ausstattung der Arbeitsstätte sowie ein kompetitives Gehaltspaket. Außerdem gilt es die fachlichen Aspekte der Landarzttätigkeit zu berücksichtigen. Hierbei sollte für die Fachrichtungen der Landärzte (z.B. Allgemeinmedizin, Innere Medizin) die gleichen Weiterbildungsmöglichkeiten zur Verfügung stehen und die Ausstattung der Arbeitsstätte dem entsprechen, was einem angemessenen Stand der Fachrichtung entspricht. Während Ersteres unter anderem durch Anpassungen der bestehenden Abläufe möglich ist, z.B. durch die Einführung von Onlineweiterbildungen, so wird für Letzteres eher finanzielle Unterstützung benötigt. Hierbei existieren bereits einige Programme auf verschiedenen Ebenen, so dass auch dies in die nationale Strategie zur Bekämpfung des Landärztemangels eingebettet werden kann.

Weiterhin gilt es, die nötige Infrastruktur zu schaffen. Diese geht über die medizinische Infrastruktur hinaus und umfasst auch die Bedürfnisse des sozialen Umfelds der Landärzte. So ist zuallererst eine akzeptable Anbindung an städtische Gebiete unabdingbar. Daneben werden Betreuungseinrichtungen für Kinder und Arbeitsmöglichkeiten für Partner vor Ort benötigt. Weiterhin gilt es sicherzustellen, dass eine Weiterbildung für den Landarzt möglich ist. Insbesondere hierbei, aber auch für die Bedürfnisse von Partner und Kinder, kann dies in den Ausbau der digitalen Infrastruktur in Deutschland mit eingebunden werden. Somit lassen sich die Maßnahmen zur Bekämpfung des Landärztemangels in einen größeren Rahmen einordnen. Es existieren bereits zahlreiche Initiativen, wie z.B. das Bundesprogramm „Ländliche Entwicklung“ des Bundesministeriums für Ernährung und Landwirtschaft, welches die Zukunftsfähigkeit des ländlichen Deutschlands steigern sollen (Bundesministerium für Ernährung und Landwirtschaft, 2019), so dass auf bestehende Erfolge aufgebaut werden kann.

6. Limitationen und Ausblick für die Forschung

Diese Arbeit unterliegt einer Reihe an Limitationen, die sich in methodische und inhaltliche Limitationen einteilen lassen. Diese sollten in zukünftigen Studien adressiert werden. Zusätzlich konnten einige Richtungen für potentielle weitere Studien identifiziert werden.

Auf methodischer Ebene finden sich Limitationen aufgrund der Nutzung eines DCEs mit MNL, der Berechnung der WTA sowie der Stichprobe selber. Das größte Problem bei der Nutzung eines DCEs ist im Fall dieser Arbeit, dass die Probanden nicht tatsächlich die Bereitschaft zeigen müssen, sich zu verpflichten. Es wird zwar davon ausgegangen, dass das echte Verhalten der Probanden ähnliche Eigenschaften aufzeigt, jedoch können im Falle der Handlungsempfehlungen Abweichungen zur Stichprobe auftreten.

Zusätzlich zeigt die MNL weitere Einschränkungen. Dazu gehören auf Seiten der Messung der Präferenzen Kontexteffekte und sogenanntes „Extreme Response Behavior“. So können Kontexteffekte aufgrund der vorhandenen No-Choice-Option aufgetreten sein. Wenn die beiden Alternativen gleich attraktiv sind, kann es sein, dass die No-Choice-Option als Kompromiss gewählt wird. Andere Kontexteffekte werden von Schlereth and Skiera (2017) diskutiert und beschrieben. In einigen Fällen ist auch „Extreme Response Behavior“ aufgetreten, bei dem Probanden niemals oder immer die No-Choice-Option wählen. Auf der Seite der Modellierung nimmt MNL die Homoskedastizität sowie voneinander unabhängige Fehlerterme, von denen weiterführend auf die Unabhängigkeit von irrelevanten Alternativen („independence of irrelevant alternatives“) geschlossen werden kann, an. Für weiterführende Studien gilt es prüfen, ob diese Annahmen nicht zu lockern sind und andere Modellierungen zu verwenden sind. Ryan et al. (2008) geben hier ein Überblick über weiteren Modellierungen und Schlereth and Skiera (2017) über neuartigere Möglichkeiten im Bereich der DCE.

Des Weiteren ist die Beschreibung der Willingness-to-Accept als negative Willingness-to-Pay (siehe Unterkapitel 3.4) nicht vollkommen passend. So tritt aufgrund der Schätzung mit MNL auch häufig ein negativer Preisparameter auf, welcher eine negative WTA erzeugt. Wie vorhergehend beschrieben, ist dies nicht logisch, so dass daher die negativen WTAs gleich null gesetzt wurden. Das nachfolgende Histogramm zeigt beispielhaft die Verteilung der WTAs für das Programm der KV Brandenburg für die Anstellung im Krankenhaus. Es wird deutlich sichtbar, dass der Durchschnitt von 1470 € im Monat weit vom Median mit 237 € im Monat entfernt ist. Dies könnte durch andere Modellierungen der WTA, wie z.B. unter Zuhilfenahme einer lognormalen Verteilung, verbessert werden.

So stellt auch die Stichprobengröße von 335 Probanden eine gewisse Limitation dar. Zwar liegt sie, wie in Unterkapitel 3.3 beschrieben, deutlich über verschiedenen Mindeststichprobengrößen, die in einigen methodischen Veröffentlichungen diskutiert wurden (de Bekker-Grob et al., 2015). Nichtsdestotrotz konnte in der Stichprobe nicht über

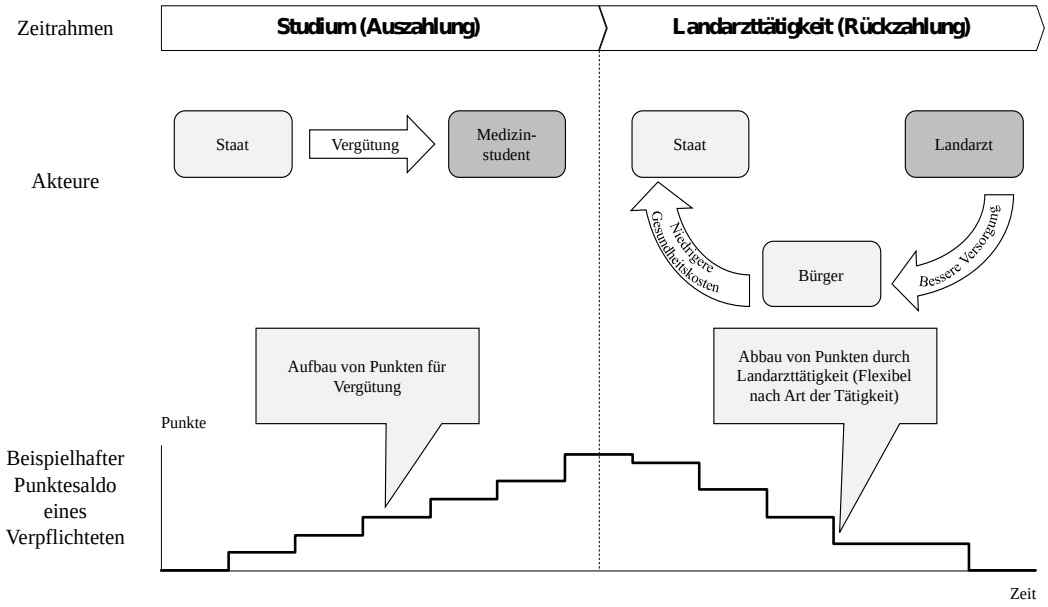


Figure 7: Überblick des "Umgekehrten Rentensystems"

Quelle: Eigene Abbildung

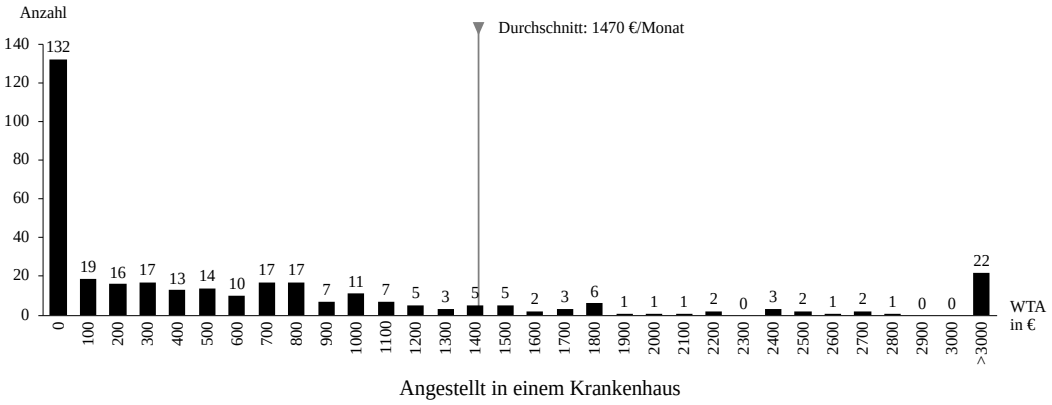


Figure 8: Histogramm der WTA für Landärzteprogramm der KV Brandenburg

Quelle: Eigene Darstellung, Landärztinnen/Landärzte-Richtlinie (2019)

mehr als ein Merkmal differenziert werden, da ansonsten die Zahl der Probanden in der Kategorie zu klein geworden wäre, um eine Aussage zu treffen. Damit konnten Aussagen, die auf der Verbindung von Merkmalen beruhen, wie z.B. Medizinstudenten unter 22 Jahren, die vom Land kommen, nicht getroffen werden.

Zusätzlich stellt die Stichprobe zwar einen heterogenen Ausschnitt aus der Gesamtbevölkerung der Medizinstudenten in Deutschland dar, ist jedoch nicht repräsentativ. Obwohl Probanden von allen medizinischen Fakultäten, mit Ausnahme der Medizinischen Hochschule Brandenburg vertreten waren, stechen insbesondere die Ruhr-Universität Bochum, die Johann-Wolfgang-von-Goethe-Universität Frankfurt am Main, die Universität des Saarlandes und die Universität zu Köln heraus. Von den medizinischen Fakultäten dieser vier Universitäten kamen über ein Drittel der Probanden der Umfrage. Zusätzlich wurde die Umfrage während der COVID-19-Pandemie und der damit verbundenen Kontaktsperre in Deutschland durchgeführt. So kann die Stichprobenzusammensetzung dadurch beeinflusst worden sein, dass einige Teile der Studentenschaft in der medizinischen Versorgung ausgeholfen haben (Richter-Kuhlmann, 2020).

Auf der inhaltlichen Seite zeigen Limitationen auch weitere Möglichkeiten für die zukünftige Forschung auf. Der Ort der Anstellung der Verpflichtung wurde hier über Einwohnerzahl und Erreichbarkeit des nächsten Ballungsgebiets definiert. Dies deckt jedoch nicht die Fülle der Möglichkeiten ab, mit denen ein Ort charakterisiert werden kann. Weitere Kriterien könnten die Zusammensetzung der Infrastruktur oder die bauliche Struktur beinhalten Marschner et al., 2018.

Weiterführende Studien sollten ein ganzheitlicheres Konzept zur Beschreibung eines Ortes entwickeln, dieses auf Arbeitsorte eines Landarztes anwenden und dann die Präferenzen dazu erkunden.

Weiterhin zeigte sich in den Parameterschätzungen, dass Variationen häufig mit demografischen Merkmalen, wie Alter, Studienabschnitt oder Beziehungsstatus, korrelieren. Dies lässt vermuten, dass dahinter eine latente Variable, wie „Lebensalter“ oder „Lebenserfahrung“, liegt, welche die Präferenzen maßgeblich beeinflusst. So sollten zukünftige Studien diese latente Variable weiteruntersuchen, weitere Indikatoren finden und anhand dieser die Variation in den Präferenzen besser erklären können.

Zusätzlich gilt es auch die Gründe hinter dem negativen Preisparameter zu verstehen. Eine mögliche Fragestellung, die hierbei untersucht werden sollte, ist das Vorliegen des sogenannten „Mental Accounting Bias“, bei dem der Nutzwert des Geldes nicht von der Menge, sondern anderen Kriterien, wie z.B. der Quelle, abhängig ist (Thaler, 1999). Daher könnte es möglich sein, dass bei einigen Probanden der Nutzwert des Geldes geringer ist, da das Geld in Verbindung mit einer Einschränkung der Freiheit durch die Verpflichtung wahrgenommen wird. Denkbar wären hier andere Ansätze der Modellierung sowie angepasste Formen des DCEs, wie zuvor diskutiert, um dies zu umgehen.

7. Zusammenfassung

Diese Arbeit hat sich zum Ziel gesetzt, das Problem des Landärztemangels in Deutschland zielorientiert zu untersuchen. Hierbei ging es speziell darum, in wieweit Anreizsysteme für Medizinstudenten strukturiert werden können, damit diese nach ihrer Approbation in ländlichen Gebieten arbeiten, um dem dort vorherrschenden Ärztemangel entgegenzuwirken. In ihrer Funktion als medizinische Grundversorger sind Landärzte ausschlaggebend für die Gesundheit der Bevölkerung auf dem Land, so dass ein Mangel an diesen die Ungleichheit in der Gesellschaft erhöht und ein hohes Maß an vermeidbaren medizinischen Folgekosten verursacht (Guanais & Macinko, 2009; Macinko et al., 2010). Anstelle Substitutionslösungen dieses Problems zu untersuchen, fokussiert sich diese Arbeit auf Anreize, die Medizinstudenten geboten werden müssen, damit diese als Landarzt praktizieren. Um dem zukunftsbezogenen Aspekt Genüge zu tun, hat sich diese Arbeit ausschließlich auf die aktuellen Medizinstudenten und damit auf die zukünftige Generation an Ärzten konzentriert.

Methodisch wurde dies im Rahmen eines Discrete-Choice-Experiments untersucht. Dabei wurde den Probanden ein Programm angeboten, bei dem sie sich gegen eine Vergütung verpflichten können, nach dem Studium für eine bestimmte Zeit auf dem Land zu arbeiten. Im Rahmen weiterer Eigenschaften des DCEs wurden auch Präferenzen bezüglich Arbeitsstätten, Teilzeitarbeit und Unterstützung bei der Facharztweiterbildung analysiert.

Die Ergebnisse des DCEs zeigen nicht nur, dass die Bereitschaft einer Verpflichtung auf dem Land zu arbeiten bei einem Teil der Medizinstudenten in Deutschland vorhanden ist, sondern auch, welche Untergruppen dafür besonders empfänglich sind. Vor allem gilt es, ein mögliches Verpflichtungsprogramm flexibel in Bezug auf Dauer, Ort und genaue Modalitäten zu gestalten, da aufgrund der Heterogenität der Medizinstudenten kein Universalprogramm entworfen werden kann. So gilt jedoch, dass erstens, ein Verpflichtungsprogramm für all jene Medizinstudenten attraktiv ist, die sich eher am Anfang des Studiums befinden und auf dem Land aufgewachsen sind. Weiterhin stehen insbesondere alleinstehende oder weibliche Medizinstudenten dem Verpflichtungsprogramm positiv gegenüber . Zweitens bewegt sich der Rahmen der monatlichen Zahlungen hierbei von rechnerischen Nullbeträgen bis hin zum ungefähren monatlichen Bedarf eines Studenten. Drittens kann darüber hinaus die Verpflichtung durch die Wahl der bevorzugten Arbeitsstätte, der Möglichkeit, Teilzeit zu arbeiten, sowie durch Unterstützung bei der Facharztweiterbildung attraktiver gestaltet werden. Viertens wird im qualitativen Teil der Analyse besonders die genaue Wahl des Ortes der Verpflichtung hervorgehoben, da oftmals bestimmte Regionen Deutschlands als Arbeitsort gewünscht werden. Fünftens ist die Erfüllung weiterer infrastruktureller Bedürfnisse, wie z.B. Arbeitsstelle für den Partner, Betreuung für Kinder, Anbindung an das nächste Ballungsgebiet, für viele potentielle Landärzte wichtig.

Anhand dessen wird im Rahmen der Handlungsempfehlung ein sogenanntes „Umgekehrtes Rentensystem“ entwickelt. Dieses stellt ein Verpflichtungsprogramm dar, bei dem sich, wie im DCE, Studenten für die Arbeit als Landarzt gegen eine monatliche Vergütung verpflichten. Die Höhe der Vergütung sowie der Ort und die Länge der Verpflichtung werden dabei über ein Punktesystem in Relation gesetzt, um die von den Studenten benötigte Flexibilität zu erreichen. Dieses Programm sollte deutschlandweit von Bund und Ländern gemeinsam eingeführt werden, um das Problem eines Landärztemangels in Deutschland zu bekämpfen. Zusätzlich wird dieses Programm als Investment des Staates in die Gesundheit des Bürgers herausgestellt, wobei durch die Verbesserung der Grundversorgung Folgekosten reduziert werden können. Folglich kann der Staat seine grundgesetzliche Pflicht, dem Bürger ein funktionierendes Gesundheitssystem zur Verfügung zu stellen, erfüllen. Die genaue Berechnung dieser finanziellen Einsparungen für den Staat wird anhand einiger Fakten angenommen, jedoch wird die endgültige Beweisführung zukünftigen Studien überlassen.

Ferner werden weitere Anknüpfungspunkte für zukünftige Studien beschrieben. Diese gehen sowohl auf inhaltliche Aspekte, bei denen das Thema dieser Arbeit fortgeführt werden würde, sowie methodische Aspekte, bei denen insbesondere die Modellierung modifiziert werden würde, ein.

References

Amt für Statistik Berlin-Brandenburg. (2020). *Studierende an Hochschulen im Land Brandenburg (endgültige Angaben)*. Retrieved from https://www.statistik-berlin-brandenburg.de/publikationen/stat_berichte/2018/SB_B03-01-00_2017j01_BB.pdf

Arrow, K. J. (1963). Uncertainty and the Welfare Economics of Medical Care. *The American Economic Review*, *53*(5), 941–973.

Bauer, C. (2017). *Landarztquote: Irrweg oder gute Idee?* Retrieved from https://www.bayerisches-aerzteblatt.de/fileadmin/aerzteblatt/ausgaben/2017/05/einzelpdf/BAB_5_2017_227.pdf

Blumenthal, D. M., Culpepper, L., Stafford, R., & Starfield, B. (1999). The Duration of Ambulatory Visits to Physicians. *Journal of Family Practice*, *48*(4), 261–271.

Bodenheimer, T., & Pham, H. (2010). Primary care: current problems and proposed solutions. *Health Affairs*, *29*(5), 799–805.

Bolduc, D., Laval, U., & Fortin, B. (1996). The Effect of Incentive Policies on the Practice Location of Doctors. *Journal of Labor Economics*, *14*(4), 703–732.

Bremer, P., & Wübker, A. (2013). Sozioökonomische Unterschiede in der Inanspruchnahme von Haus- und Facharztleistungen in Deutschland: Eine empirische Analyse. *Prävention und Gesundheitsförderung*, *8*(1), 15–21.

Bärnighausen, T., & Bloom, D. E. (2009). Financial Incentives for Return of Service in Underserved Areas: A Systematic Review. *BMC Health Services Research*, *9*(1), 86–103.

Bundesministerium der Finanzen. (2020). *Bundeshaushalt.de: Einzelplan 2020, Soll - Ausgabe #1501 - Gesetzliche Krankenversicherung.* Retrieved from https://www.bundeshaushalt.de/#/2020/soll/ausgaben/einzelplan/1501.html

Bundesministerium für Ernährung und Landwirtschaft. (2019). *Bundesprogramm "Ländliche Entwicklung" – Ideen und Impulse für die Zukunft unserer ländlichen Räume.* Retrieved from https://www.bmel.de/SharedDocs/Downloads/DE/Broschueren/bundesprogramm-laendliche-entwicklung2019.pdf

Bundesministerium für Gesundheit. (2020). *Bundesländer.* Retrieved from https://www.bundesgesundheitsministerium.de/themen/gesundheitswesen/staatliche-ordnung/bundeslaender.html

Chen, L. M., Farwell, W. R., & Jha, A. K. (2009). Primary care visit duration and quality: Does good care take longer? *Archives of Internal Medicine*, *169*(20), 1866–1872.

Childress, J. E., Faden, R. R., Gaare, R. D., Gosm, L., Kahn, J., Bonnie, R. J., … Nieburg, P. (2002). Public Health Ethics: Mapping the Terrain. *The Journal of Law, Medicine & Ethics*, *30*(2), 170–178.

Clark, M. D., Determann, D., Petrou, S., Moro, D., & de Bekker-Grob, E. W. (2014). Discrete Choice Experiments in Health Economics: A Review of the Literature. *PharmacoEconomics*, *32*(9), 883–902.

Constitution of the World Health Organization. (2006). *Basic Documents* . Retrieved from https://www.who.int/governance/eb/who_constitution_en.pdf?ua=1

Cooper, C. L., Rout, U., & Faragher, B. (1989). Mental Health, Job Satisfaction, and Job Stress Among General Practitioners. *BMJ*, *289*(6670), 366–370.

Cutchin, M. P. (1997a). Physician Retention in Rural Communities: The Perspective of Experiential Place Integration. *Science & Medicine*, *44*(11), 1661–1674.

Cutchin, M. P. (1997b). Physician Retention in Rural Communities: The Perspective of Experiential Place Integration. *Health & Place*, *3*(1), 25–41.

Cutler, D. M., & Zeckhauser, R. J. (2000). *The Anatomy of Health Insurance.*

Daniels, Z. M., Van Leit, B. J., Skipper, B. J., Sanders, M. L., & Rhyne, R. L. (2007). Factors in Recruiting and Retaining Health Professionals for Rural Practice. *The Journal of Rural Health*, *23*(1), 62–71.

de Bekker-Grob, E. W., Donkers, B., Jonker, M. F., & Stolk, E. A. (2015). Sample Size Requirements for Discrete-Choice Experiments in Healthcare: A Practical Guide. *The Patient-Patient-Centered Outcomes Research*, *8*(5), 373–384.

Deutscher Bundestag. (2017). *Einundzwanzigster Bericht nach § 35 des Bundesausbildungsförderungsgesetzes zur Überprüfung der Bedarfssätze, Freibeträge sowie Vomhundertsätze und Höchstbeträge nach § 21 Absatz 2.* Retrieved from http://dipbt.bundestag.de/doc/btd/19/002/1900275.pdf

de Vise, P. (1973). Physician Migration from Inland to Coastal States: Antipodal Examples of Illinois and California. *Academic Medicine*, *48*(2), 141–151.

Dionne, G., Langlois, A., & Lemire, N. (1987). More on the Geographical Distribution of Physicians. *Journal of Health Economics*, *6*(4), 365–374.

Doescher, M. P., Ellsbury, K. E., & Hart, L. G. (2000). The Distribution of Rural Female Generalist Physicians in the United States. *The Journal of Rural Health*, *16*(2), 111–118.

Donabedian, A. (1966). Evaluating the Quality of Medical Care. *The Milbank Memorial Fund Quarterly*, *44*(3), 166–206.

Dubben, H. H., Muche-Borowski, C., & Scherer, M. (2016). Kann man Über- und Unterversorgung messen? *Zeitschrift Für Allgemeinmedizin*, *92*(10), 414–418.

Dussault, G., & Franceschini, M. C. (2006). Not enough there, too many here: Understanding geographical imbalances in the distribution of the health workforce. *Human Resources for Health*, *4*(1), 12–28.

el Koussa, M., Atun, R., Bowser, D., & Kruk, M. E. (2016). Factors influencing physicians' choice of workplace: Systematic review of drivers of attrition and policy interventions to address them. *Journal of Global Health*, *6*(2), 1–13.

Europäische Zentralbank. (2020). *Bilateral Exchange Rates Time Series*. ECB Statistics Bulletin. Retrieved from http://sdw.ecb.europa.eu/servlet/desis?node=10000071

Fruen, M. A., Hadley, J., & Korper, S. P. (1980). Effects of financial incentives on physicians' speciality and location. *Health Policy and Educatlon*, *1*(2), 143–159.

Geyman, J., Hart, L. G., Norris, C. J. B., T. E., & Lishner, D. M. (2000). Education for rural practice educating generalist physicians for rural practice: How are we doing? *The Journal of Rural Health*, *16*(1), 56–80.

Günther, O. H., Kürstein, B., Riedel-Heller, S. G., & König, H. H. (2010). The role of monetary and nonmonetary incentives on the choice of practice establishment: A stated preference study of young physicians in Germany. *Health Services Research*, *45*(1), 212–229.

Green, P. E., & Srinivasan, V. (1978). Conjoint analysis in consumer research: Issues and outlook. *Journal of Consumer Research*, *5*(2), 103–123.

Grembowski, D., Paschane, D., Diehr, P., Katon, W., Martin, D., & Patrick, D. L. (2005). Managed care, physician job satisfaction, and the quality of primary care. *Journal of General Internal Medicine*, *20*(3), 271–277.

Guanais, F., & Macinko, J. (2009). Primary Care and Avoidable Hospitalizations: evidence From Brazil. *Journal of Ambulatory Care Manage*, *32*(2), 115–122.

Han, G. S., & Humphreys, J. S. (2006). Integration and retention of international medical graduates in rural communities: A typological analysis. *Journal of Sociology*, *42*(2), 189–207.

Hancock, C., Steinbach, A., Nesbitt, T. S., Adler, S. R., & Auerswald, C. L. (2009). Why doctors choose small towns: A developmental model of rural physician recruitment and retention. *Social Science and Medicine*, *69*(9), 1368–1376.

Hartmannbund. (2019). *Landarztquote ist kein Zukunftskonzept*. Retrieved from https://www.hartmannbund.de/detailansichten/landesverbaende/niedersachsen/presse/meldung/landarztquote-ist-kein-zukunftskonzept-reimann-sollte-im-bundesrat-eine-breite-und-moderne-analy/

Hartmannbund. (2020). *Wo die Landarztquote kommt …* . Retrieved from https://www.hartmannbund.de/berufspolitik/informationen/wo-die-landarztquote-kommt/

Hegney, D., Mccarthy, A., Rogers-Clark, C., & Gorman, D. (2002). Retaining rural and remote area nurses: The Queensland, Australia experience. *JONA: The Journal of Nursing Administration*, *32*(3), 128–135.

Henger, R., & Oberst, C. (2019). *Alterung der Gesellschaft im Stadt-Land Vergleich.* Retrieved from https://www.iwkoeln.de/fileadmin/user_upload/Studien/Kurzberichte/PDF/2019/IW-Kurzbericht_2019_Alterung_der_Gesellschaft.pdf

Herd, M. S., Bulsara, M. K., Jones, M. P., & Mak, D. B. (2017). Preferred

practice location at medical school commencement strongly determines graduates' rural preferences and work locations. *Australian Journal of Rural Health*, *25*(1), 15–21.

Heszen-Klemens, I., & Lapińska, E. (1984). Doctor-patient interaction, patients' health behavior and effects of treatment. *Social Science & Medicine*, *19*(1), 9–18.

Holst, J., Normann, O., & Herrmann, M. (2015). Strengthening training in rural practice in Germany: New approach for undergraduate medical curriculum towards sustaining rural health care. *Rural and Remote Health*, *15*.

Johnson, R., & Orme, B. (2003). *Getting the most from CBC* (Tech. Rep.). Sequim: Sawtooth Software Technical Paper. Retrieved from https://sawtoothsoftware.com/resources/technical-papers/getting-the-most-from-cbc-2003

Joyce, C. M., McNeil, J. J., & Stoelwinder, J. U. (2004). Time for a new approach to medical workforce planning. *The Medical Journal of Australia*, *180*, 343–346.

Kasch, R., Engelhardt, M., Förch, M., Merk, H., Walcher, F., & Fröhlich, S. (2016). Ärztemangel: Was tun, bevor Generation Y ausbleibt? Ergebnisse einer bundesweiten Befragung. *Zentralblatt Für Chirurgie - Zeitschrift Für Allgemeine, Viszeral- Und Gefäßchirurgie*, *141*(2), 190–196.

KBV. (2016). *Deutschlandweite Projektion 2030 – Arztzahlentwicklung in Deutschland.* Retrieved from https://www.kbv.de/media/sp/2016_10_05_Projektion_2030_Arztzahlentwicklung.pdf

KBV. (2018a). *Berufsmonitoring Medizinstudierende 2018.* Retrieved from https://www.kbv.de/media/sp/Berufsmonitoring_Medizinstudierende_2018.pdf

KBV. (2018b). *KBV Gesundheitsdaten - Versorgungsgrade nach Planungsregion.*

KBV. (2020). *Reform der Bedarfsplanung.* Retrieved from https://www.kbv.de/media/sp/Instrumente_Bedarfsplanung_Broschuere.pdf

Kistemann, T., & Schröer, M. A. (2007). Kleinräumige kassenärztliche Versorgung und subjektives Standortwahlverhalten von Vertragsärzten in einem überversorgten Planungsgebiet. *Gesundheitswesen*, *69*(11), 593–600.

Kolstad, J. R. (2011). How to make rural jobs more attractive to health workers. Findings from a discrete choice experiment in Tanzania. *Health Economics*, *20*(2), 196–211.

Kristiansen, I. S., & Førde, O. H. (1992). Medical specialists' choice of location: The role of geographical attachment in Norway. *Social Science & Medicine*, *34*(1), 57–62.

KV Baden-Württemberg. (2018). *Weiterbildungsförderung.* Retrieved from https://www.kvbawue.de/praxis/niederlassung/foerderung-informationsangebot/weiterbildungsfoerderung/

KV Brandenburg. (2019). *Stipendium für Studierende der Medizin.* Retrieved from https://www.kvbb.de/praxis/studium-weiterbildung/foerderprogramm/stipendium/

KV Niedersachsen. (2017). KVN und Land starten Stipendium-Programm zur Landärzte-Gewinnung. *KVN pro: Informationen der Kassenärztlichen Vereinigung Niedersachsen*, *12*(47), 1–1.

Lamu, A. N., & Olsen, J. A. (2016). The relative importance of health, income and social relations for subjective well-being: An integrative analysis. *Social Science & Medicin*, *152*, 176–185.

Landesgesetz zur Sicherstellung der hausärztlichen Versorgung in ländlichen Regionen des Landes Rheinland-Pfalz (LArztG RP). (2019).

Landkreis Waldeck-Frankenberg. (2017). *Landarzt werden.* Retrieved from https://www.landarzt-werden.de/landkreis-waldeck-frankenberg.html

Landärztinnen/Landärzte-Richtlinie. (2019). *Förderrichtlinie des Ministeriums für Arbeit, Soziales, Gesundheit, Frauen und Familie des Landes Brandenburg zur Stärkung der landärztlichen Versorgung Brandenburgs*.

Laven, G., & Wilkinson, D. (2003). Rural doctors and rural backgrounds: How strong is the evidence? A systematic review. *The Australian Journal of Rural Health*, *11*(6), 277–284.

Luppa, M., Giersdorf, J., Riedel-Heller, S., Prütz, F., & Rommel, A. (2020). Frequent attenders in the German healthcare system: Determinants of high utilization of primary care services. Results from the cross-sectional German health interview and examination survey for adults (DEGS). *BMC Family Practice*, *21*(1), 1–10.

Macinko, J., Dourado, I., Aquino, R., de Fátima Bonolo, P., Lima-Costa, M. F., Medina, M. G., … Turci, M. A. (2010). Major expansion of primary care in Brazil linked to decline in unnecessary hospitalization. *Health Affairs*, *29*(12), 2149–2160.

Macinko, J., Starfield, B., & Shi, L. (2007). The contribution of primary care systems to health outcomes within Organization for Economic Cooperation and Development (OECD) countries, 1970-1998. *International Journal of Health Services*, *37*(1), 111–126.

Maier, C. B., & Aiken, L. H. (2016). Task shifting from physicians to nurses in primary care in 39 countries: A cross-country comparative study. *European Journal of Public Health*, *26*(6), 927–934.

Marschner, C., Wasem, J., & Heymann, R. (2018). Delineating concept for defining the country practitioner – An objective approach for implementation in the field of health services research. *Gesundheitswesen*, *80*(6), 564–572.

Marschner, C., Wasem, J., & Heymann, R. (2019). The rural doctor's selection algorithm (La-Abal): Approach to implementation in the field of health care research. *Gesundheitsökonomie und Qualitätsmanagement*, *24*(3), 126–140.

Maslow, A. H. (1954). *Motivation and personality*. Harper & Brothers.

Matsumoto, M., Okayama, M., Inoue, K., & Kajii, E. (2005). Factors associated with rural doctors' intention to continue a rural career: a survey of 3072 doctors in Japan. *The Australian Journal of Rural Health*, *13*(4), 219–225.

Mayo, E., & Mathews, M. (2006). Spousal perspectives on factors influencing recruitment and retention of rural family physicians. *Canadian Journal of Rural Medicine*, *11*(4), 271–276.

McCormick, M. C. (1985). The contribution of low birth weight to infant mortality and childhood morbidity. *New England Journal of Medicine*, *312*(2), 82–90.

McFadden, D. (1974). *Conditional logit analysis of qualitative choice behavior* (P. Zarembka, Ed.). Amsterdam University Press.

Medical Tribune. (2019). Bei der Landarztquote in NRW zählt die Berufserfahrung. *Medical Tribune*. Retrieved from https://www.medical-tribune.de/meinung-und-dialog/artikel/bei-der-landarztquote-in-nrw-zaehlt-die-berufserfahrungnbsp/

Middendorff, E., Apolinarski, B., Becker, K., Bornkessel, P., Brandt, T., Heißenberg, S., & Poskowsky, J. (2017a). *Die wirtschaftliche und soziale Lage der Studierenden in Deutschland 2016 – Zusammenfassung zur 21. Sozialerhebung des Deutschen Studentenwerks durchgeführt vom Deutschen Zentrum für Hochschul- und Wissenschaftsforschung.* Retrieved from https://www.bmbf.de/upload_filestore/pub/21._Sozialerhebung_2016_Zusammenfassung.pdf

Middendorff, E., Apolinarski, B., Becker, K., Bornkessel, P., Brandt, T., Heißenberg, S., & Poskowsky, J. (2017b). *Die wirtschaftliche und soziale Lage der Studierenden in Deutschland 2016 – 21. Sozialerhebung des Deutschen Studentenwerks durchgeführt vom Deutschen Zentrum für Hochschul- und Wissenschaftsforschung.* Retrieved from http://www.sozialerhebung.de/download/21/Soz21_hauptbericht.pdf

Mokkink, R. G. H., Smits, A., van Eijk, J., Beek, M., Mesker, P., & Mesker-Niesten, J. (1985). Work satisfaction of general practitioners and the quality of patient care. *Family Practice*, *2*(3), 128–135.

Oberlandesgericht Düsseldorf. (2020). *Düsseldorfer Tabelle.* Retrieved from https://www.olg-duesseldorf.nrw.de/infos/Duesseldorfer_Tabelle/Tabelle-2020/Duesseldorfer-Tabelle-2020.pdf

OECD. (2016). Graph 5.1 - Physician density, by Territorial Level 2 regions, 2013 (or nearest year). *Health Workforce Policies in OECD Countries: Right Jobs, Right Skills, Right Places, OECD Health Policy Studies*.

Ono, T., Schoenstein, M., & Buchan, J. (2014). *Geographic imbalances in doctor supply and policy responses.* Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract{_}id=1008326

Orme, B. (1996). *Sample size issues for conjoint analysis studies* (Tech. Rep.). Sequim: Sawtooth Software Technical Paper. Retrieved from https://sawtoothsoftware.com/resources/technical-papers/sample-size-issues-for-conjoint

-analysis-studies-2019

Parchman, M. L., & Culler, S. (1994). Primary care physicians and avoidable hospitalizations. *Journal of Family Practice*, *39*(2), 123–128.

Parsons, T. (1967). Definition von Gesundheit und Krankheit im Lichte der Wertbegriffe und der sozialen Struktur Amerikas. In A. Mitscherlich (Ed.), *Der Kranke in der modernen Gesellschaft* (pp. 57–87). Kiepenheuer & Witsch.

Pathman, D. E., Konrad, T. R., & Agnew, C. R. (1994). Studying the retention of rural physicians. *The Journal of Rural Health*, *10*(3), 183–192.

Pereira, L. L., Santos, L. M. P., Santos, W., Oliveira, A., & Rattner, D. (2016). Mais Médicos program: provision of medical doctors in rural, remote and socially vulnerable areas of Brazil, 2013-2014. *Rural and Remote Health*, *16*(1).

Pohontsch, N. J., Hansen, H., Schäfer, I., & Scherer, M. (2018). General practitioners' perception of being a doctor in urban vs. rural regions in Germany - A focus group study. *Family Practice*, *35*(2), 209–215.

Porter, M. E., & Guth, C. (2012). *Chancen für das deutsche Gesundheitssystem: Von Partikularinteressen zu mehr Patientennutzen*. Springer-Verlag.

PricewaterhouseCoopers. (2018). *Das deutsche Gesundheitswesen auf dem Prüfstand.* Retrieved from https://www.pwc.de/de/gesundheitswesen-und-pharma/pwc-das-deutsche-gesundheitswesen-auf-dem-pruefstand.pdf

Rau, R., & Schmertmann, C. P. (2020). District-level life expectancy in Germany. *Deutsches Ärzteblatt Int*, *117*, 493–499.

Richter-Kuhlmann, E. (2020). Medizinstudium während der Pandemie: Suche nach Lösungen. *Deutsches Ärzteblatt*, *117*(13), A-645-B-550.

Robertson, J. E., Boyd, J., Hedges, J. R., & Keenan, E. J. (2007). Strategies for increasing the physician workforce: The Oregon model for expansion. *Academic Medicine*, *82*(12), 1158–1162.

Roick, C., Heider, D., Günther, O. H., Kürstein, B., Riedel-Heller, S. G., & König, H. H. (2012). Was ist künftigen Hausärzten bei der Niederlassungsentscheidung wichtig? Ergebnisse einer postalischen Befragung junger Ärzte in Deutschland. *Gesundheitswesen*, *74*(1), 12–20.

Rosenblatt, R. A., Whitcomb, M. E., Cullen, T. J., Lishner, D. M., & Hart, L. (1992). Which medical schools produce rural physicians? *JAMA*, *268*(12), 1559–1565.

Rosenthal, M. B., Zaslavsky, A., & Newhouse, J. P. (2005). The geographic distribution of physicians revisited. *Health Services Research*, *40*(6I), 1931–1952.

Rowe, I. S., C. J. andCampbell, & Hargrave, L. A. (2014). Rural experience for junior doctors: Is it time to make it mandatory? *The Australian Journal of Rural Health*, *22*(2), 63–67.

Ryan, M., Gerard, K., & Amaya-Amaya, M. (2008). Using Discrete Choice Experiments to Value Health and Health Care. In M. Ryan, K. Gerard, & M. Amaya-Amaya (Eds.), (Vol. 11). Springer Netherlands.

Schlereth, C., Eckert, C., & Skiera, B. (2012). Using discrete choice experiments to estimate willingness-to-pay intervals. *Marketing Letters*, *23*(3), 761–776.

Schlereth, C., & Skiera, B. (2017). Two new features in discrete choice experiments to improve willingness-to-pay estimation that result in SDR and SADR: Separated (Adaptive) Dual Response. *Management Science*, *63*(3), 829–842.

Schlereth, C., Skiera, B., & Schulz, F. (2018). Why do consumers prefer static instead of dynamic pricing plans? An empirical study for a better understanding of the low preferences for time-variant pricing plans. *European Journal of Operational Research*, *269*(3), 1165–1179.

Schmidt, M. (2008). Das Trauerspiel um die Weiterbildung zum Facharzt für Allgemeinmedizin - Persönliche Bestandsaufnahme aus der Sicht eines Weiterbildungsassistenten. *Zeitschrift für Allgemeinmedizin*, *84*(1), 10–13.

Scott, A., Witt, J., Humphreys, J., Joyce, C., Kalb, G., Jeon, S.-H., & Mcgrail, M. (2013). Getting doctors into the bush: General Practitioners' preferences for rural location. *Social Science & Medicine*, *96*(1), 33–44.

Shi, L., Macinko, J., Starfield, B., Xu, J., Regan, J., Politzer, R., & Wulu, J. (2004). Primary care, infant mortality, and low birth weight in the states of the USA. *Journal of Epidemiology and Community Health*, *85*(5), 374–380.

Shi, L., & Starfield, B. (2000). Primary care, income inequality, and self-rated health in the United States: A mixed-level analysis. *International Journal of Health Services*, *30*(3), 541–555.

Shi, L., Starfield, B., Kennedy, B., & Kawachi, I. (1999). Income inequality, primary care, and health indicators. *Journal of Family Practioners*, *48*(4), 275–284.

Süptitz, J., & Schlereth, C. (2017). Fracking: Measuring social acceptance and the effect of acceptance increasing measures. *Schmalenbachs Zeitschrift für Betriebswirtschaftliche Forschung*, *69*(4), 405–439.

Starfield, B., Shi, L., & Macinko, J. (2005). Contribution of primary care to health systems and health. *The Milbank Quarterly*, *83*(3), 457–502.

Statistisches Bundesamt. (2016). *Older People in Germany and the EU.* Retrieved from https://www.destatis.de/EN/Themes/Society-Environment/Population/Current-Population/Publications/Downloads-Current-Population/brochure-older-people-eu-0010021169004.pdf

Statistisches Bundesamt. (2019). *Gesundheitsausgaben im Jahr 2018 um 4 % gestiegen.* Retrieved from https://www.destatis.de/DE/Presse/Pressemitteilungen/2020/05/PD20_164_23611.html

Statistisches Bundesamt. (2020). *Konsumausgaben privater Haushalte, Fachserie 15, Heft 5 - 2018.* Retrieved from https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Einkommen-Konsum-Lebensbedingungen/Konsumausgaben-Lebenshaltungskosten/Publikationen/Downloads-Konsumausgaben/evs-aufwendung-privater-haushalte-2152605189004.pdf?__blob=publicationFile

Sundmacher, L., & Ozegowski, S. (2016). Regional distribution of physicians: The role of comprehensive private health insurance in Germany. *European Journal of Health Economics*, *17*(4), 443–451.

Swait, J., & Andrews, R. L. (2003). Enriching scanner panel models with choice experiments. *Marketing Science*, *22*(4), 442–460.

Szafran, O., Crutcher, R. A., & Chaytors, R. G. (2001). Location of family medicine graduates' practices: What factors influence Albertans' choices? *Canadian Family Physician*, *47*(11), 2279–2285.

Thaler, R. H. (1999). Mental accounting matters. *Journal of Behavioral Decision Making*, *12*(3), 183–206.

Train, K. E. (2009). *Discrete Choice Methods with Simulation*. Cambridge University Press.

Ubach, C., Scott, A., French, F., Awramenko, M., & Needham, G. (2003). What do hospital consultants value about their jobs? A discrete choice experiment. *BMJ*, *326*(7404), 1432–1438.

van den Bussche, H., Schwartz, F. W., & Angerer, P. (2010). Arbeitsbelastung und Berufszufriedenheit bei niedergelassenen Ärztinnen und Ärzten: Genug Zeit für die Patientenversorgung. *Arbeitsbedingungen und Befinden von Ärztinnen und Ärzten*, *2*, 235–242.

verdi. (2018). *Entgelttabelle Pflege- und Betreuungseinrichtungen.* Retrieved from https://www.verdi.de/file5c9b844c0596fb4f8ccdb126/download/Entgelttabellen-VKA-Pflege.pdf

Walendzik, A., Manouguian, M., Greß, S., & Wasem, J. (2009). Vergütungsunterschiede im ambulanten ärztlichen Bereich zwischen PKV und GKV und Modelle der Vergütungsangleichung. *Sozialer Fortschritt*, *58*(4), 63–69.

Whitlock, E. P., Orleans, C. T., Pender, N., & Allan, J. (2002). Evaluating primary care behavioral counseling interventions: An evidence-based approach. *American Journal of Preventive Medicine*, *22*(4), 267–284.

Wibulpolprasert, S., & Pengpaibon, P. (2003). Human resources for health integrated strategies to tackle the inequitable distribution of doctors in Thailand: four decades of experience. *Human Resources for Health*, *1*(1), 12-29.

Wissenschaftlicher Dienst des Deutschen Bundestages. (2015). *Grundgesetzlicher Anspruch auf gesundheitliche Versorgung*.

Witt, J., Scott, A., & Osborne, R. H. (2009). Designing choice experiments with many attributes. An application to setting priorities for orthopaedic waiting lists. *Health Economics*, *18*(6), 681–696.

Yong, J., Scott, A., Gravelle, H., Sivey, P., & McGrail, M. (2018). Do rural incentives payments affect entries and exits of general practitioners? *Social Science and Medicine*, *214*(1), 197–205.

Zeckhauser, R. (1970). Medical insurance: A case study of the tradeoff between risk spreading and appropriate incentives. *Journal of Economic Theory*, *2*(1), 10–26.

Junior Management Science 6(3) (2021) 547-567

Junior Management Science

journal homepage: www.jums.academy

The Impact of Community Involvement on Game Life-Cycle: Evidence based on Gaming Platform Steam

Tim Leonard Hermes

Ludwig-Maximilians-Universität München

Abstract

Later stages of the product life-cycle are characterized by diminishing sales and declining prices. Especially firms with substantial product development costs, as is the case in the video game industry, are dependent on long product life-cycles to amortize initial costs. This confronts firms with the fundamental challenge of maintaining the value of their product from the consumer's perspective and thus delaying the natural price decline. We investigate whether product features that facilitate community involvement and interaction are an effective means to keep the product stimulating and relevant in the long run. Using extensive data from the PC video game market, we show that the inclusion of interactive, community-engaging features allows firms to both charge higher prices and delay the natural price decline of their product. However, for one of the investigated features we find the opposite effect, which we explain by subsequent analysis. Thereby, we gain valuable insights into the importance of robustly designed incentive systems in community-focused features. Our findings could help firms in their efforts to design attractive and economically viable products with prolonged life-cycles.

Keywords: Product life-cycle; digital goods pricing; user communities; co-creation; digital gaming platforms.

1. Introduction

Every product is subject to a life-cycle (Levitt, 1965). This applies to digital products just as much as to physical products. Although they do not age through wear and tear, they fall victim to psychological obsolescence (Park, 2010). Later stages of the product life-cycle are characterized by diminishing sales and declining prices (Golder & Tellis, 2003). From an economic point of view, a product's life-cycle ends when it no longer generates revenue. While progressive digitization of content substantially reduced marginal cost, there are still many digital products whose consumption is preceded by significant fixed costs, be it apps, movies, music or video games (Shy, 2001, p. 182). This constitutes the highly challenging task for firms to ensure long life-cycles and thereby amortize initial costs.

Extending the life-cycle translates into remaining relevant from the customer's perspective by offering continuous high value, which in turn allows firms to delay the natural price decline of their product. While traditional quality features of a product, such as the graphic quality of a video game, inevitably become obsolete with age, firms are increasingly turning their attention to product features that generate long-term stimulation and relevance. In the age of connectedness, it is the interactive, community-involving features that drive ongoing engagement (Von Hippel & Katz, 2002). In theory, such "Community Features" empower users to interact, set their own impulses, exchange ideas and give feedback. Thus, the community becomes a living "sounding board" which keeps the product vibrant and dynamic in the long run. Thereby, firms hope to maintain the value of the product from the consumer's perspective over the product life-cycle, and to capture it through higher prices.

In this context we attempt to address the **research question** of how Community Involvement influences price setting of firms on digital platforms? We investigate this question using the PC video game market as a relevant example, for four reasons. First, from an economic standpoint, video gaming has matured into a worldwide market worth 130B USD with an annual growth rate of 5.8% over the last five years, surpassing other two-sided markets such as music, cinema, TV or radio (PwC, 2019). Further, it is predicted that within the next five years consumers will spend an average of 25% more time annually on video games, leaving the growth rates

DOI: https://doi.org/10.5282/jums/v6i3pp547-567

of any other digital platforms like music streaming and social networking behind (Clarefield, 2018; PwC, 2017). Second, from a practical standpoint, we argue that long life-cycles are of particular relevance here. Unlike many other types of digital content, consumption is preceded by years of expensive development. Bankruptcies of developers who were unable to amortize their game during its lifespan are the pervasive reality (Leibovitz, 2013). Third, digital gaming platforms share many characteristics of efficient markets, most notably, many suppliers and consumers, low barriers to entry, fairly rational consumers, intense competition, high transparency and independent pricing by suppliers (see Chapter 3: Industry Background). This makes them an intriguing object of investigation as a representative for other digital platforms. Lastly, from a market dynamics standpoint, the video game industry is currently undergoing an upheaval that challenges the previous understanding of the term "gaming platform". Previously, platforms in the video game context were mostly understood as physical goods in the form of consoles, which act as an interface between game developers and players. Recently, through digitization of content, advanced digital rights management systems and higher performance of personal computers, we have witnessed the rise of "digital gaming platforms", allowing developers to publish PC games cheaply and consumers to buy and play them directly on the platform (Jöckel, Will, & Schwarzer, 2008). These new digital gaming platforms, which act as a central hub for millions of players and thousands of developers, allow developers to involve their community in entirely new ways. We are focusing on the pioneering and largest platform "Steam".

For this research we are investigating three novel methods that game developers can employ to promote player interaction and involvement. Specifically, we look at three Community Features that allow players to (1) customize a game with own and peers' User Generated Content, (2) collect and trade virtual items and (3) complete game challenges and earn social status, respectively. We develop theory to argue that the inclusion of these Community Features in games allows developers to charge higher prices and delay the natural price decline over the life-cycle of their game.

To test this empirically, we collected five years worth of longitudinal data on pricing, game characteristics and player numbers for over 2,000 games on Steam. For Community Feature (1) and (3) we find that developers can both charge higher prices and delay the natural price decline. For feature (2) we find the opposite effect. We explain the unexpected latter result with a subsequent analysis. We discover that the feature was used in a different way than the platform Steam intended and point out the importance of designing robust incentive systems during the conception of Community Features. We thereby contribute to the sparse empirical research on the effects of Community Involvement on pricing in the context of digital platforms.

The rest of this paper is structured as follows. Chapter 2 presents an overview of the related literature on pricing, life-cycles and Community Involvement. Chapter 3 provides a closer look of digital gaming platforms and Steam. Chapter 4 develops theory and hypotheses. Chapter 5 describes methods used for data collecting and analysis, ventures an exploratory descriptive analysis and features our models. Chapter 6 presents results, while Chapter 7 discusses them and shows limitations. Finally, Chapter 8 concludes with implications and directions for future research.

2. Related Literature

Our study touches on a number of key theoretical concepts, mainly pricing, product life-cycle and Community Involvement in the context of digital platforms. We will give a brief overview of how these topics relate in previous literature.

2.1. Pricing and Product Life-Cycle

Literature on pricing of new products is vast and covers various aspects. In this chapter we provide a selective overview of pricing decisions over the life-cycle of a product. Two prominent pricing policies in particular have been much discussed in the literature: A strategy of high introductory prices that fall over time to skim off the demand, "price skimming", and the strategy of low initial prices to capture the market, "penetration pricing" (Chatterjee, 2009; Dean, 1976). We will mainly focus on skimming strategies. Also relevant is the abstract notion of product lifecycles. While definitions are often fuzzy, it most commonly describes the gaussian-like diffusion in form of sales of a product over time along with declining prices (Levitt, 1965).

Early studies show the effects pricing can have on the diffusion, i.e. sales, of a new product. Already Bass and Bultez (1982) found, using theoretical diffusion models, that prices should fall monotonously over time. Later, Krishnan, Bass, and Jain (1999) specified that a monotonically falling price is optimal, especially if the new product is confronted with a price-sensitive target group or if the use of the new product involves risk, e.g. lack of quality. This is in line with similar studies using diffusion models (Bass, 1980; Dolan & Jeuland, 1981; Kalish, 1983). Unlike previously discussed studies that assume monopolistic market situations, a more recent study in the context of oligopolistic market structures suggests that optimal prices drop to a lower level if the product is subject to a shorter life-cycle, i.e. speed of diffusion is higher (Dockner & Fruchter, 2004). The authors justify this by arguing that producers try to gain larger market shares in the face of a shortened product life through aggressive pricing.

These early studies have a few things in common: first, the optimal price is solved analytically in a theoretical model, second, the logic is not tested on real world data, and most importantly third, demand is assumed to be exogenous (i.e. given via diffusion mechanism) and undynamic (in terms of strategic response to producer pricing). Motivated by these deficits, Nair (2004) shows, based on data from the video games market, that over the life of a product, producers are confronted with consumers that have different willingness to pay. Those who attach high value to the product buy it

early, while those with a low willingness to pay wait. According to Nair, companies thus have an incentive to reduce prices as the product ages to skim off individual willingness to pay with so-called intertemporal price discrimination. Interestingly, he further shows that consumer behavior can be rational and strategic by anticipating price reductions and postponing the purchase decisions. Foundational literature supports the notion of such heterogenous demand conditions and attributes them to varying consumer preferences, which translate into different anticipated utility and willingness to pay (Roberts & Lilien, 1993; Varian, 1992). Rietveld and Eggers (2018) show for the video game market that games encounter a heterogeneous audience over their life-cycle, starting with early adopters who are willing to experiment and have significantly higher willingness-to-pay and late adopters who are more risk-averse.

However, while all these studies recognize and model that consumers are heterogenous, they also base their analyses on the questionable assumption that perceived utility of a product is constant over time for each individual (Adner & Levinthal, 2001). Although preferences can be considered mostly static, products are not "fixed", Adner and Levinthal point out. Utility and therefore willingness to pay can be increased over time by continuously introducing more quality to the product (Adner & Levinthal, 2001). Unlike most other studies they deviate from the simplifying assumption of monopolistic market situations and show that constant "rejuvenation" is especially effective with increasing competition. Similarly, Weerahandi and Dalal (1992) adapt the usual diffusion models to include the effect of increased utility through product attributes on purchase decisions. In contrast to most aforementioned research, a more recent study deviates from the assumption of exogenous demand (i.e. diffusion mechanism), by forecasting sales of a product purely based on increased product utility over time (Fruchter & Orbach, 2008). They even go a step further as they also consider characteristics that constitute positive network externalities. Their findings suggest that prices can be held up if product characteristics are improved.

The previously discussed literature argues that continually falling prices are the optimal pricing policy. This is also observable in studies of the video game market (Liu, 2010; Nair, 2004) and in our video game data, too. However, we would like to briefly mention that penetration pricing, i.e. initially aggressive prices, can be observed in other contexts. For example, in the platform battle between Playstation and Nintendo, Liu (2010) shows that initial penetration pricing can foster essential indirect network effects between console and game developers in order to win the market.

From the previously considered studies, some overarching insights can be deduced: First, over the lifespan of a product, the optimal prices fall. Second, much of the price research looks at the impact of pricing (independent variable) on sales (dependent variable). Thus, price is usefully seen as a strategic marketing tool. In the following, we want to reverse the role of price and consider it as a measure of value for a product that should be maximized over the life-cycle. Third, there is little research on how the inclusion of value-generating product features can increase the willingness to pay over time and allow higher prices.

2.2. Community Involvement as Utility Dimension

There is diverse literature that investigates the influence of product characteristics on consumer utility and willingness to pay. For example, previous studies have investigated changes in utility and willingness to pay for changes in product characteristics for digital goods such as music downloads (Hahsler & Breidert, 2006), smartphones (Böhm, Adam, & Farrell, 2015), mobile service platforms (Nikou & et al., 2014), online banking services (Yusuf Dauda & Lee, 2015) and tablet news apps (Gundlach & Hofmann, 2017).

However, in this study we are interested in the effect of Community Features. We define Community Features as product features that involve the community and enable interactions between platform participants. They are particularly relevant in the context of platforms as they are used in interaction with other users and can hence exhibit considerable network utility, more discussed in Chapter 4: Theory and Hypotheses (Katz & Shapiro, 1985; Meyer, 2012). Nonetheless, we find that the influence of Community Features on utility is very rarely measured. As one of few, a study by Benlian, Vetter, and Hess (2010) shows that Community Features (e.g. playlist sharing) in music streaming have a positive effect on the willingness to pay, but less so than classic quality features (e.g. sound quality). A comparable study for online games shows that willingsness to pay is higher for Community Features (e.g. virtual chat rooms) and customization features (e.g. character creation) than for better quality features (e.g. graphic quality). This indicates that Community Involvement is valued differently depending on the product. Yet, both analyses only provide a static view on the value of these features. Two very recent studies on the video game market adopt a more dynamic view by showing that Community Features can have a lasting effect on the life-cycle of a product. First, a study of the video game market found that games with multiplayer functionality generate higher sales in later phases of the life-cycle (Marchand, 2016). Second, Rietveld and Eggers (2018) show for Free-To-Play Games on Steam that games with Community Features (online multiplayer) are much more likely to become a "killer app", measured by download numbers.

While multiplayer functionality in video games is a well-established way of letting individuals interact with each other and benefit from direct network effects, there are many other ways to involve the community that haven't been subject to much prior empirical research.

3. Industry Background

Traditionally, video gaming platforms have mainly been understood as physical goods in the form of consoles, which act as an interface between game developers and players. Thus, the video game market can be classified as a two-sided

market, which facilitates interaction between complementors, i.e. producers of complementary products (game developers), and consumers (players) (Eisenmann, Parker, & Van Alstyne, 2006). The video game market is characterized by a number of features, most notably: indirect network effects, which suggest that the utility obtained by platform participants on one side depends on the number of complementary platform participants on the other side (Katz & Shapiro, 1985). Equally important are direct network effects, which state that the utility also depends on the number of participants on the same platform side (Shy, 2011). In addition, the traditional video game market is characterized by intense hardware standards battles (Gallagher & Park, 2002), varying inter-generational hardware compatibility (Claussen, Kretschmer, & Spengler, 2010), short and interdependent hardware and software life-cycles (Marchand, 2016), the phenomenon of multi-homing between platforms (Landsman, V. and Stremersch, S., 2011), heterogeneous consumer needs (Rietveld & Eggers, 2018) and the presence of high switching costs (Gallagher & Park, 2002).

However, the video games industry is currently undergoing a transformation that challenges the traditional understanding of the term "games platform". With the digitalization of content, advanced Digital Rights Management systems and increased performance of computers, a new intermediary has entered the market: so-called "Digital Game Distribution Platforms" allow developers to offer PC games at low cost and players to buy and play games directly on the platform (Jöckel et al., 2008). The pioneer and largest representative of such digital gaming platforms is Steam, with 40,000 games and 20M active players at peak times (Steamspy, 2020).

While these digital gaming platforms introduce many novelties, the central mechanics of the traditional gaming industry, such as network effects and game lifecycles, are preserved. Still, some rules are changing. First, there is hardly any dependency between hardware generations and games anymore, since the compatibility of PC games is not dependent on a specific hardware generation, as was usually the case with consoles (Marchand, 2016). This means that the life-cycle of games is not predefined by the life-cycle of hardware. Second, there are low barriers to entry, which gave rise to the genre of "indie" games (Lin, Bezemer, & Hassan, 2018). Small and semi-professional developers ("indie developers") are able to publish their game for a one-time payment of 100€ and get exposure to an audience of millions, whereas before only big software companies could afford development and were exploited by platform owners as the "money-making side" (Bowen Martin & Deuze, 2009; Steam, 2020d). Third, digital platforms are characterized by increased transparency, as prices and independent ratings are directly comparable, which was previously less feasible with analogue games. Fourth, Nair (2004) has already shown for the traditional video game industry that consumers can act rationally and strategically by anticipating price reductions and postponing purchases. It is likely that such rational consumer behavior increases in the face of greater market transparency on digital gaming platforms.

Given these attributes, digital gaming platforms share many characteristics of efficient markets; most notably, many suppliers and consumers, fairly rational consumers, low barriers to entry, intense competition, high transparency and independent pricing by suppliers (Bork, 1993). This makes Steam an intriguing object of investigation as a representative for other digital platforms.

Finally, these new digital gaming platforms, which act as a central hub for millions of players and thousands of developers, enable entirely new possibilities for interaction and Community Involvement. For example, in recent years Steam has developed a range of value-added services that game developers have been able to integrate voluntarily through a Software Development Kit. These features have been targeted at further involving players, or as Steam puts it in their developer documentation, "to increase engagement, encourage player interaction and motivate players to spend more of their time in-game" (Steam, 2020c). This is a prime example of a platform owner, aspiring to empower its complementors with powerful tools to sustain long-term utility of their content.

4. Theory and Hypotheses

In this study we investigate the influence of Community Involvement on the price-setting of companies. In the context of the PC video game industry, we look at three important ways in which developers let their players get involved and interact with each other. In the following, we take a closer look at these three ways and derive testable hypotheses. "Community Feature" is not an established term and is rarely used in the literature. For our study, we broadly define those product features as Community Features that engage users by enabling interactions between platform members. Platform members are both users and game developers.

4.1. Contributing User Generated Content: "Steam Workshop"

Giving users the opportunity to participate in product development, create complementary products, and modify existing products is not a new phenomenon. Chesbrough (2003) coined the term "open innovation" to describe how companies can better meet the needs of their target group by complementing their own innovation efforts with external user innovation. According to Urban and Von Hippel (1988), so-called lead users, i.e. users who feel the needs of the target group much earlier than the rest of the market, are of particular importance. Their contributions to the innovation process are considered more relevant and novel by other users (Lilien & et al., 2001; Schreier, 2007; Von Hippel, 1986).

User innovation has a long tradition in the video game market. Games have long sought to differentiate themselves in the face of intense competition and to maximize customer satisfaction by giving them a say in company decisions (Jeppesen, 2005; Koch & Bierbamer, 2016). For example,

the success of bestseller "The Sims" is often explained by the richness of User Generated Content (e.g. game characters), which is estimated to account for up to 90%. (CERI, 2006; Banks, 2013). A case study of the game "Trackmania" concludes that the game's unusually long life span can be explained by the fact that the developers allow users to continually expand the game, which has led to a self-sustaining ecosystem of over 100 community websites and 3 million active hobby developers (Koch & Bierbamer, 2016; Parmentier & Gandia, 2013).

Since 2011, Steam offers a marketplace called *Workshop*. Developers can offer a *Workshop* for their game, allowing users to create, share and integrate content such as characters, game items and levels. Currently, there are about 4,000 games using this functionality, compared to 200 in 2016. In addition, developers provide user toolkits, i.e. tools and guidelines that allow users to create new products, complementary products or modify existing ones (Von Hippel & Katz, 2002). In the context of Steam, developers provide their users with so-called Software Development Kits, a collection of game-specific developer tools that allow players to create compatible content for the game at hand. A positive side effect of *Workshop* is that the developer facilitates a community that acts like a living "sounding board". It allows players to share best practices, allows mutual help and learning, which increases the value of the game (Humphreys, 2005; Jeppesen, 2005). It also creates bilateral communication and feedback mechanisms between developer and player.

Jeppesen (2004) argues that the value-added through the availability of User Generated Content is twofold. First, the game becomes more attractive through the fact that complementary, often free content is available. The many possibilities for "mix and match", according to Jeppesen, keep the game young in the long term, which has a positive effect on sales, especially in later phases of life. Second, and less intuitively, the author shows that developers systematically integrate popular User Generated Content into their game. This suggests that such games continuously increase their value from the user's perspective and are less likely to be out of touch with the needs of the target group. We found proof of this phenomenon in Steam's Subscription Agreement: "In some cases, *Workshop* Contributions may be considered for incorporation by Valve or a third-party developer into a game" (Steam, 2020b).

Nevertheless, an extensive case-study, which conducted interviews with game industry executives and experts, reveals that the share of innovating players is low, whereas the vast majority are not willing "to improve or modify or exchange with others" (Burger-Helmchen & Cohendet, 2011). Further, interviews with these regular players indicate that content from others is occasionally integrated and may lead to higher satisfaction and loyalty. However, previous research has shown that the quality of User Generated Content varies greatly (Lastowka, 2007). Thus, it is possible that users have bad experiences due to bugs, incompatibility, malware or offensive content (e.g. pornography), which negatively rubs off on the value of the underlying game. Steam counteracts this with rating systems, Steam moderators and the possibility for developers to curate contributions beforehand. However, it is unclear how many developers use this, given the effort involved.

When studying the motivation behind User Generated Content, we must distinguish between extrinsic and intrinsic factors. Extrinsic motivators include monetary incentives. For example, Steam lets users charge money for their creations to "allow mod [modification] makers the opportunity to work on their mods full time" (Steam Blog, 2015, 2020). This encountered great resistance from players, as user content has traditionally been free, which is why nowadays only a fraction of all content has to be paid for. According to Lerner and Triole (2002), building a reputation within the community is more important for users. In some cases, hobby developers even manage to break into developer jobs (Cadin, Guérin, & Defillippi, 2006; Poor, 2014). Nevertheless, intrinsic motivators seem to be the main driver for user innovation. A recent study, which conducted a survey of user innovators, identified joy, learning new skills and cognitive challenges as key (Koch & Kerschbaum, 2014). Franke and Von Hippel (2002) also point out the heterogeneity of user needs that cannot be met by standard software products. Users must therefore accept products that do not fit them well, which reduces their perceived value of the game. By offering toolkits, users get a voice and can tailor the product to their needs, whereby lead users benefit directly as they bring the necessary skills and resources, and other users indirectly by using the shared content (Franke & Von Hippel, 2002; Koch & Bierbamer, 2016).

Koch and Bierbamer (2016) are able to show for a small sample of games that the ability to create User Generated Content has a positive influence on the rating of a game. We investigate whether this higher satisfaction can be exploited with higher prices and therefore formulate the hypotheses:

> **H1a** *Enabling players to create own game content and enjoy their peers' content via Workshop allows developers to charge higher prices.*

> **H1b** *Enabling players to create own game content and enjoy their peers' content via Workshop allows developers to delay the price decline of their game.*

We know that the full effect of User Generated Content is achieved when it not only helps one individual to customize their game, but is also available to all others, whereby everyone benefits in form of network utility (Katz & Shapiro, 1985). It can therefore be assumed that a larger community will benefit more from the possibility of creating User Generated Content. We therefore formulate:

> **H1c** *A larger player community allows developers to charge higher prices for enabling players to create and share content.*

4.2. Collecting and Trading Virtual Items: "Steam Tradingcards"

Virtual items that players can earn, collect and trade have become an integral part of online games. They can take many different forms, such as in-game items, trophies or trading cards, and can entail diverse functionalities, benefits and values. Since 2013, Steam has been offering developers the opportunity to create game-specific *Tradingcards*. Some of the cards are given to players at random intervals during the game, while the remaining cards must be acquired by trading with other players on Steam (or buying them with virtual Steam money) to get a complete set. This set can then be displayed on the profile or exchanged for tradable benefits such as badges, personalized Steam profile backgrounds and game coupons (Steam, 2020a). About 9,000 games offer *Tradingcards*, which is about 20% of all games (Steam Card Exchange, 2020).

A large number of studies have examined the psychological stimuli triggered by virtual goods as well as the needs these goods satisfy. Bostan (2009) argues that players satisfy materialistic needs by possessing virtual goods, just as they do in the physical world. Some players spend substantial amounts of time amassing items to show them off, compare and trade them (Guo & Barnes, 2007; Yee, 2006). Such collectors often feel a strong desire for completeness (Danet & Katriel, 1994). Steam takes advantage of this by requiring players to have a full set of cards to reap the full benefits. This "un-locking mechanism" evokes anticipation and player interest in the long-term (Hallford & Lamothe, 2001; Wang & Sun, 2011). It is also conceivable that less mainstream games will become disproportionately more attractive through *Tradingcards*. For collectors, their cards are particularly valuable due to their rarity (Hamari & Lehdonvirta, 2010).

Formanek (1994) argues that collectibles also have a sentimental value as they remind of successful gaming experiences. In the case of Steam, *Tradingcards* can be a personal expression allowing a player to publicly identify with a game, signal one's interests and be part of a certain gaming community (Wang & Sun, 2011).

In the case of Steam, the incentives go beyond the mere possession of items. Players can use their cards to customize profile backgrounds, generate badges, and use rare emoticons while chatting, thereby "cosmetically" differentiating themselves. While this satisfies intrinsic needs for customization (Guo & Barnes, 2007), it also creates a "visual authority" over others, which many players appreciate, according to a survey of 300 players (Park & Lee, 2011).

Getting a full deck of cards is only possible by trading with others. Thereby, Steam adds a social dimension to a game. Given that social interaction is one of the most important gratifications of gaming, a positive influence on the value of a game can be expected, especially for single-player games (Scharkow, 2015). However, it is possible that rewards for activities outside the game, i.e. trading, or through mere playing time are perceived as unfair. Steam counteracts this by offering rewards that are only "cosmetic" and not functional, thereby persevering performance-oriented game mechanics.

Offering rewards at random times during the game has a gambling character, which makes such games exciting and addictive (Wang & Sun, 2011). Steam takes advantage of this logic by distributing cards at unknown times. However, in extreme cases, according to Rozin (2000), this can have negative effects on the gaming experience if the time of the reward is not predictable at all.

Virtual items can also be extrinsically motivating. Castronova (2005, p. 19) exemplifies the value virtual assets can have by showing that the multiplayer game Everquest is amongst the 80 largest economies in the world and its currency is more valuable than the Yen. Whang and Kim (2005) use a survey of nearly 5,000 players to show that some players see collectibles as a pure investment with value appreciation. Rare cards on Steam also embody enormous monetary value. They can be exchanged for virtual money on Steam, exchanged for real money on eBay or an enormous ecosystem at dedicated websites (e.g. steamcardexchange.net), as well as converted into expensive games.

Based on intrinsic and extrinsic motivators, we formulate the following hypotheses:

> **H2a** *Enabling players to collect and trade virtual items in form of Tradingcards allows developers to charge higher prices.*
>
> **H2b** *Enabling players to collect and trade virtual items in form of Tradingcards allows developers to delay the price decline of their game.*

4.3. Mastering Challenges and Earning Status: "Steam Achievements"

There is a wide variety of reward-based systems that target intrinsic and extrinsic motivations to motivate individuals towards high performance, commitment and achievement of goals (Farzan & et al., 2008). The influence of rewards such as virtual badges, points or rankings has previously been investigated by empirical studies for a variety of areas, for example enterprise social networks (Farzan & et al., 2008; Thom, Millen, & DiMicco, 2012), collaborative ideation tools (Jung, Schneider, & Valacich, 2010), e-learning tools (Denny, 2013), peer lending platforms (Hamari, 2013) and image tagging platforms (Mekler, Brühlmann, Opwis, & Tuch, 2013).

The games industry is considered the origin of such reward systems, hence the term "gamification" (Hamari, 2013). Since 2007, games on Steam have been able to reward players for particular accomplishments in the game with so-called *Achievements*. Developers define challenges of different difficulty, some of which are closely interwoven with the story, and some of which are mundane tasks. Upon completion, players will receive *Achievements*, essentially badges, which will be displayed on their Steam profile.

Based on a survey of 3,000 players, Yee (2006) identifies the feeling of achievement, which is expressed in advance-

ment and competition, as one of the most important motivational factors in video games.

Literature suggests that rewards, such as *Achievements*, create value for players by setting clear goals (Farzan & et al., 2008; Hamari, 2013). This is in line with goal-setting theory, according to which individuals are motivated to high performance by challenging, short-term, measurable and specific goals (Locke & Latham, 1991, 2002). In addition, meaningful contextual challenges that teach players the possibilities of a game can bring considerable pleasure. Thus, *Achievements* are to be understood as a "guiding instrument" that helps the player to progress in the game (Hamari & Eranti, 2011; Montola, Nummenmaa, Lucero, Boberg, & Korhonen, 2009). Ideally, *Achievements* allow players to become completely immersed in the game and reach a mental "flow" state (Csikszentmihalyi, 1990; Wang & Sun, 2011).

Steam grants *Achievements* immediately after completing a task. According to Hamari (2013), the immediacy of positive feedback generates strong satisfaction and confirmation. This in turn has a positive influence on the perception of one's own competence, which can lead to intrinsic motivation (Mekler et al., 2013). More fundamental support for this is provided by reinforcement theory, which states that regular positive feedback strengthens the existing behavior (Luthans & Stajkovic, 1999). However, poorly designed achievement systems can cause frustration, for example, when *Achievements* are either too easy or too difficult to achieve (Juul, 2010), are used inflationarily or can be obtained through mundane tasks. In the case of Steam, where developers can define up to 500 *Achievements* per game, the quality and meaningfulness might suffer.

Nonetheless, *Achievements* are not an end in themselves. A large part of their value for individuals comes from the fact that they are publicly displayed. Thus, citetMekler2013 concluded based on a controlled experiment of letting people tag images that participants are motivated by comparability and display of performance. Similarly, Farzan and et al. (2008) demonstrates that individuals show more activity on an enterprise social network as a result of public status and points. These studies show that individuals in different contexts derive social value from features that allow comparison. This fulfills two needs central to gaming: "accumulating in-game symbols of wealth or status" and "the desire to challenge and compete with others", as Yee (2006) puts it. Barnett and Coulson (2010) note very practically that building a positive reputation and signaling competence is essential for being accepted into teams. Moreover, just as with virtual gamesitems, there are also systematic "Achievement Completists" who play a game until they have collected them all (Jakobsson, 2011, 2016). On Steam we find numerous groups like "100% Achievements Group" with thousands of members.

Finally, Hamari and Koivisto (2013) show in the context of a fitness social network that gamification features such as *Achievements* generate positive attitudes towards the product through increased reputation. This in turn leads to a higher probability of further use and significantly higher recommendation rates. Thus, *Achievements* can also be seen as an effective marketing instrument that increases the anticipated value of a game.

Based on previous theory, we formulate the following hypotheses:

> **H3a** *Enabling players to master game challenges and earn social status via Achievements allows developers to charge higher prices.*

> **H3b** *Enabling players to master game challenges and earn social status via Achievements allows developers to delay the price decline of their game.*

5. Empirical Setting

5.1. Data Sample

5.1.1. Data Collection

The data set for this study combines data from four different sources. Collecting data from different sources was necessary to build a comprehensive and, in terms of the research question, meaningful data set. As a first step, we identified the most popular game titles on Steam. We defined games as "popular" if they had at least 1,000 active players at any time since their release. This resulted in a list of 4,180 games. The reason for such a strict cut-off was the lack of data quality and availability for less popular games. Next, we obtained historical prices in USD for these games from the website "steamdb.info", which tracks prices for a variety of Steam games since November 2014. For this, we built a custom crawler with Python, which retrieved the price history for each game in our list. The purpose of the crawler was to automate an otherwise extremely time-consuming manual task, but we paid attention to slow website calls to stay well below the server's rate limit. Unsuccessful website requests, e.g. server errors, were logged and the call was repeated at a later time. The collection process took about 3.5 weeks. Ultimately, we were able to collect price histories for 3,801 games. 379 of these 3,801 games had incomplete histories because they were released before Steamdb started tracking prices, so we deleted them, resulting in a relevant sample of 2,784 games (see Figure 1). Another important part of the data set are the characteristics of games. Some of these characteristics were directly available via Steam's public developer API "api.steampowered.com", which we queried with a dedicated Python script. The other part of the game characteristics were obtained in a similar way from the website "steamspy.com", which crawls the Steam website at regular intervals. Finally, we retrieved historical player numbers for each game from the website "steamcharts.com", which documents these on a monthly basis for many Steam games since 2012.

5.1.2. Data Preparation

When obtaining the data, we made sure that all sources use official Steam Game IDs, which allowed us to easily

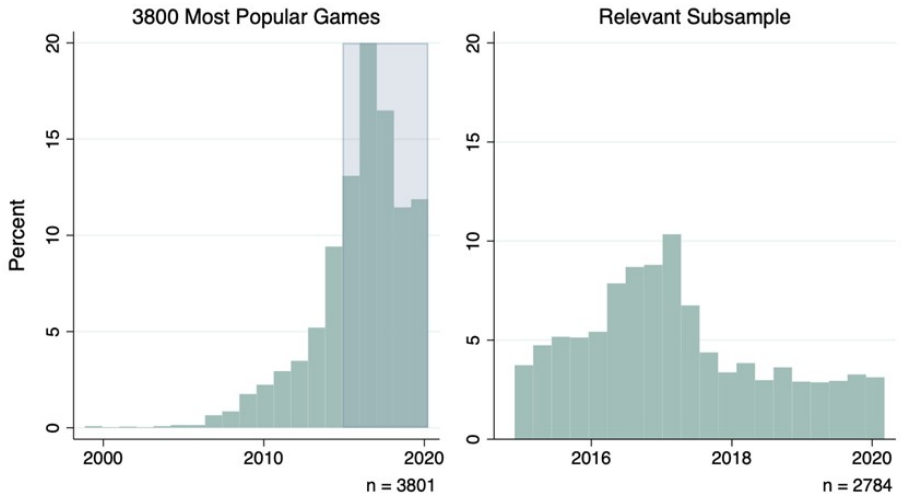


Figure 1: Distribution of Release Dates

merge the data into one large data set. From here, we identified a number of relevant cases that need to be treated differently (see Figure 2). Most importantly, we need to exclude games that have always been "free-to-play" from our sample, since a price of 0 does not reflect the value of the game, but rather indicates an alternative monetization strategy, for example via In-App-Purchases. For games that become free over time, we also assume a change in the monetization strategy and trim them from that point on. Games that retire at some point are continued with a price of 0 in the original sources, which is why we also trim them.

Game profiles on Steam are filled out manually by developers. As a result, they may fill in incomplete information when the game is released. For example, they may forget to indicate what genre the game belongs to or whether it has multiplayer functionality. Here, we use the swarm intelligence of the player base. Players can create "Tags" for each game that need to be upvoted by enough other players to appear on the profile. To increase the accuracy of variables, we linked user-defined tags to profile information.

Finally, we constructed three sub-datasets. This is necessary to answer different parts of the research question. Two are of cross-sectional structure, one of which looks at the time of release of the games, while the other compares all games at the same time-point in the present (March 2020). The third data set is panel structure and compares games over time.

5.2. Descriptive Evidence

Table 1 shows a condensed overview of the central variables across the three datasets. As we built three sub-datasets, we calculated three tables with summary statistics. Due to space restrictions, these tables can be found in the Appendix (see Table A1 for cross-sectional dataset "Release", Table A2 for cross-sectional dataset "Present", Table A3 for panel dataset).

Additionally, we take a first exploratory look at our data. Since game prices are of particular importance in our context, we will first look at the 16 (limited due to space) oldest games in our sample (see Figure 3).

While the limited number of games depicted does of course not represent all other games, we can nevertheless venture five important deductions: First, we see that in most cases the price declines over the years, which confirms that games are subject to a life-cycle. Second, we acknowledge that the rate of price decline varies greatly, some of them remaining very stable. Third, we see that price reduction is managed through sustained changes in the base price, not through discounts (represented by strong downward swings). Accordingly, the base price is a better, less volatile measure of value for a game. Fourth, we can see that games have fundamentally different price levels, e.g. some never go below $30, while others never go above $10.

Since Community Involvement is of particular importance, we also look at four features that game developers can use for this purpose. For sake of completeness, we show *Leaderboards* here, although we will limit our focus to the other three, as only they represent truly new approaches to Community Involvement. It turns out that developers integrate these features with varying frequency.

Next, we are interested to see if we can observe some indication that games with above features have different price patterns. For this purpose, we create two groups for each of the features (with and without feature) and calculate average prices for all games in the respective group for each day after release.

In fact, we can see big differences between the subgroups of each Community Feature. First, we see that the subgroups have different price levels across all features. Second, we can see that there also seem to be different price developments. Third, the feature *Tradingcards* is the only feature that performs worse than it's reference group.

The graphs should be interpreted with caution. The first impression could be misleading, as there are unobserved correlations and time effects, which we do not control at this point. However, we take this as an opportunity to take a

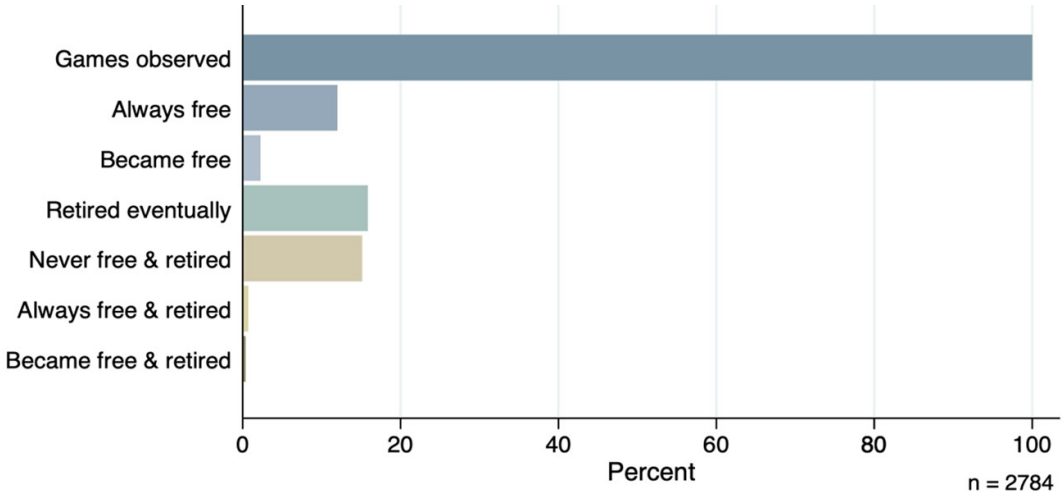


Figure 2: Categorization of noteworthy Cases

Table 1: Overview of Variables

Variable	Definition
Price	Price of a game (log)
Workshop	*User Generated Content*: Players can create game content & use others'
Tradingcards	*Virtual items*: Players can collect earn awards & display on profile
Achievements	*Challenges and Status*: Players can earn awards & display on profile
Leaderboards	Players can compare performance in rankings
Multiplayers	Game supports multiplayer
Controller Support	Game supports controllers
Cloud Gaming	Game has various cloud functionalities
TV Compatible	Game is playable on TV-Screen
Remote Gaming	Game is playable privately remotely with friends
In-app-purchase	Game offers players to buy game items
Number Languages	Number of available languages in game
Number Pictures	Number of pictures on store page to promote game
Number Trailers	Number of trailers on store page to promote game
Developer Experience	Number of games (Co-) developed so far
Publisher Experience	Number of games (Co-) published so far
Number Markets	Number of regions game is available
Platform Owner	Platform owner Valve developed the game
Genre	Game belongs to Action/Indie/Strategy/Simulation/Roleplay/ Casual/Nudity/Sport
Release Month	Release month of Games
Operating System	Game available on Windows/Mac/Linux or Combinations
User Rating	User rating for game
Number ratings	Number of total ratings for game
Average Playtime	Average individual playtime for a game to date
Number of DLC	Number of additional paid content available for game
Age	Age in days of game
Players Number	Average number of players for game over time(log)

Note: See Appendix Table A1, Table A2, Table A3 for more details & summary statistics

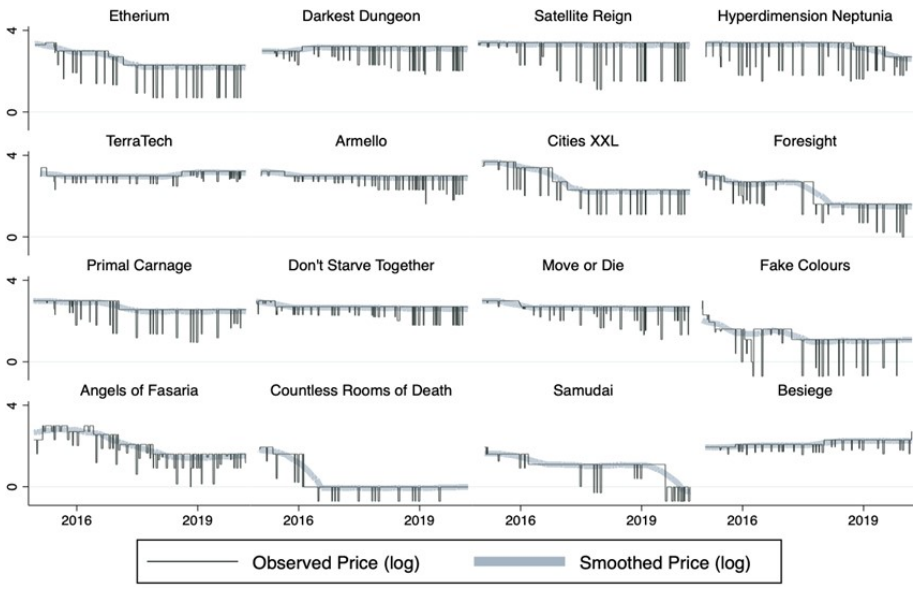


Figure 3: Price Development of earliest Games in Sample

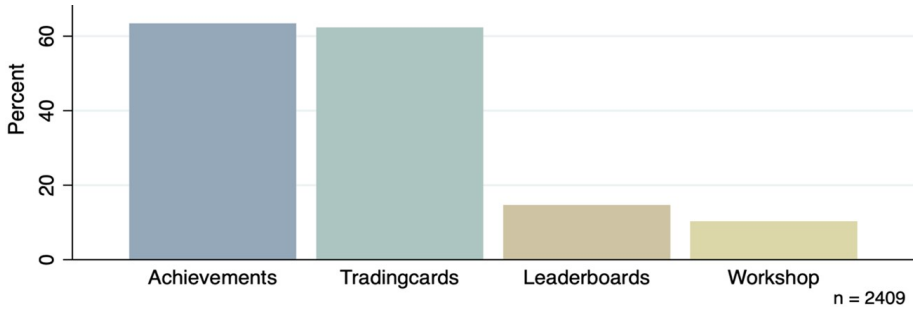


Figure 4: Proportion of Games with respective Community Feature

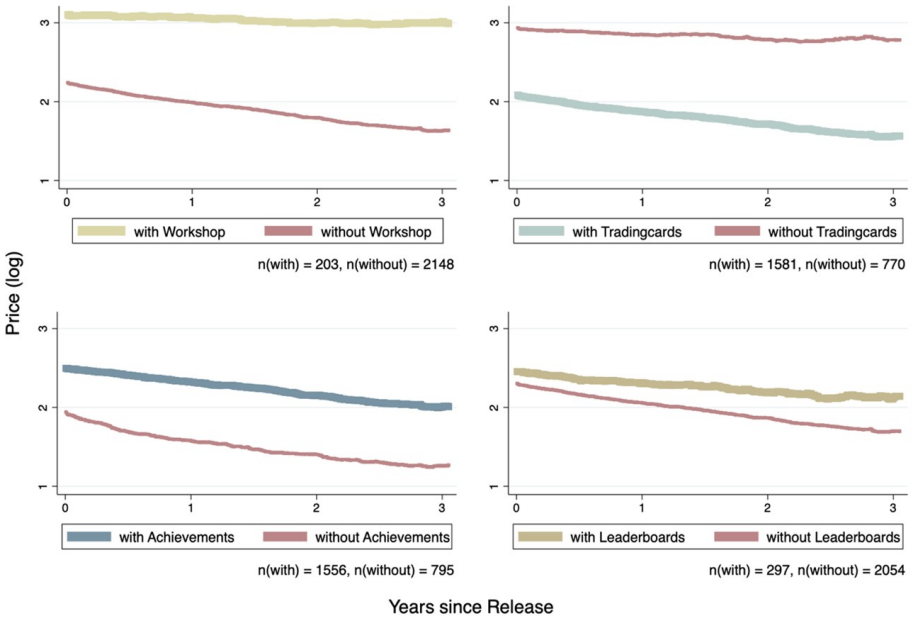


Figure 5: 3-year Price Development differentiated by Community Feature

closer look at the effect of Community Features on the lifecycle of a game.

5.3. Variables

5.3.1. Dependent Variables

The price of a game serves as a dependent variable across our analysis. Depending on the model, we either look at the price at a certain point in time, i.e. at release or present, or the price over time. Since price shows a right-skewed distribution (see Figure 6), we log-transform the price as recommended in literature (Osborne, 2010). In doing so, we strengthen the normal distribution of our Residuals (see Appendix Figure A1).

5.3.2. Independent Variables

The three Community Features, *Workshop*, *Tradingcards*, *Achievements* serve as independent variables in this study and are jointly included as Dummies. Since these three features are either implemented in a game or not, we investigate the effect of their existence on the Dependent Variable. Depending on the model, the variables are either included individually, i.e. additively, or each of them is included in interaction with a another variable, i.e. multiplicatively.

5.3.3. Control Variables

The controls we apply depend on the model. In our cross-sectional models, we employ controls to explain heterogeneity between games. Thus, we include a broad range of game characteristics that could have an impact on the quality, i.e. utility, perceived by the consumer (see Appendix Table A1 for a complete list). We also control for less tangible differences, such as increasing experience of developers and publishers over time, which could affect quality and promotion of a given game. Since we are interested in price as a value measure for game, we also control for supplementary monetization strategies such as In-App-Purchases, which might lead to a reduced price. In our panel models, we control for differences in games by including Fixed Effects, which account for all time-invariant characteristics. In addition, we control temporal factors in all models by including game age or a Dummy for each month, which accounts for both variance due to time trends and shocks. The latter is particularly important because part of our observations coincide with the outbreak of Covid-19.

5.4. Model Specification

In total, we specified four models. The first two models test H1a, H2a, H3a. We use OLS Models to exploit our data at two distinct points in time. Model 1 looks at the time of release of each game, hence we control *Release Month*.

$$\begin{aligned} Price_{iRelease} = {} & \beta_0 + \beta_1 Workshop_i + \beta_2 Tradingcards \\ & + \beta_3 Achievements_i + \beta_4 DeveloperExperience_i \\ & + \beta_5 PublisherExperience_i \\ & + \beta_6 Platformowner_i + \beta_7 NumberTrailers_i \\ & + \beta_\delta NumberPictures_i + \beta_9 NumberMarkets_i \\ & + \beta_{10} FeatureVector_i + \beta_{11} GenreVector_i \\ & + \beta_{12} OperatingsystemVector_i \\ & + \beta_{13} ReleaseMonthVector_i + \varepsilon_i \end{aligned}$$

where i denotes a specific game; $FeatureVector_i$, $GenreVector_i$, $OperatingSystemVector_i$, $ReleaseMonth\ Vector_i$ are vectors of characteristics of a game (see footnotes[1234]); ε_i represents the residuals; $Workshop_i$, $Tradingcards_i$,$Achievements_i$ are the independent Dummy variables, coded 1 if the feature exists in game and 0 otherwise.

Model 2 compares all games at the present time (March 2020), which is why we control the *Age* (non-linear, see Appendix Figure A2) to account for different lifecycle stages. Further, we include additional variables that become relevant at a later point in game life: User Rating, Number of Reviews, Average Playtime, Number of DLC.

$$\begin{aligned} Price_{i,Present} = {} & \beta_0 + \beta_1 Workshop_i + \beta_2 Tradingcards_i \\ & + \beta_3 Achievements_i + \beta_4 DeveloperExperience_i \\ & + \beta_5 PublisherExperience_i + \beta_6 PlatformOwner_i \\ & + \beta_7 NumberTrailers_i + \beta_8 NumberPictures_i \\ & + \beta_9 NumberMarkets_i + \beta_{10} UserRating_i \\ & + \beta_{12} NumberReviews_i + \beta_{12} AveragePlaytime_i \\ & + \beta_{13} NumberDLC_i + \beta_{14} GameAge_i \\ & + \beta_{15} (Age_i * Age_i) + \beta_{16} FeatureVector_i \\ & + \beta_{17} GenreVector_i \\ & + \beta_{18} OperatingSystemVector_i + \varepsilon_i \end{aligned}$$

where i denotes a specific game; $FeatureVector_i$, $GenreVector_i$, $OperatingSystemVector_i$ are vectors of characteristics of a game (same as Model 1); ε_i represents the residuals; $Workshop_i$, $Tradingcards_i$, $Achievements_i$ are the independent Dummy variables, coded 1 if the feature exists in game and 0 otherwise. Both models use more conservative Robust Standard Errors to cater for slight heteroskedasticity, which a Breusch-Pagan test revealed

[1] Vector Features of game: Leaderboards, Multiplayer, Controller Support, Cloud Gaming, Virtual Reality, TV Compatible, Remote Gaming, In-App-Purchases, Number Languages

[2] Vector Genres of game: Action, Indie, Strategy, Simulation, Roleplay, Casual, Nudity, Sport, Racing

[3] Vector Operating Systems of game: Windows, Windows&Linux, Windows&Mac, Windows&Mac&Linux

[4] Vector Release Month of games: Month Dummies for period November 2014 – March 2020

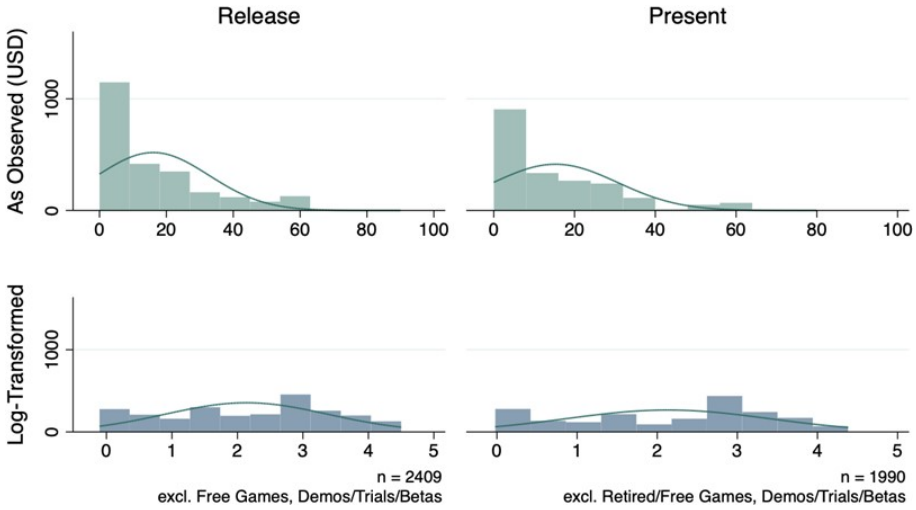


Figure 6: Distribution of Price at significant Time Points

(Breusch & Pagan, 1979). Further we calculated Variance Inflation Factors (VIFs) for our variables to ensure a lack of multicollinearity (see Appendix Table A4, Table A5). All factors are well below the commonly applied practice rule of 10 (Chatterjee & Hadi, 1986). While we do employ many variables, our variable-to-observation ratio is with 2,000 observations well below the most conservative ratio Tabachnick and Fidell (2007, p. 123) suggest.

Model 3 test H1b, H2b, H3b. Via a Hausman-Test, we tested the assumptions of applying a Panel OLS Random Effects Regression (i.e. individual game effects are uncorrelated with regressors), which allows time-invariant variables to be used in a regression (Hausman, 1978). As we couldn't reject the test-hypothesis of randomly distributed individual effects, we employ a Panel OLS Regression with Fixed Effects for games. Model 3 interacts *Age* of a game with each of the Community Features *Workshop*, *Tradingcards*, *Achievements*, which are time-invariant. Thereby, we examine whether games with these features are subject to different price development over *Age*. Further, we assume a different effect of *Player Numbers* depending on *Age*.

$$\begin{aligned} Price_{i,t} = \beta_0 &+ \beta_1(Workshop_i * Age_{i,t}) \\ &+ \beta_2(Tradingcards_i * Age_{i,t}) \\ &+ \beta_3(Achievements_i * Age_{i,t}) \\ &+ \beta_4(PlayerNumbers_{i,t} * Age_{i,t}) \\ &+ \beta_5 Age_{i,t} + \beta_6 PlayerNumbers_{i,t} + u_i + \varepsilon_{i,t} \end{aligned}$$

where i denotes a specific game, t a specific day, u_i the individual (fixed) effect of a game, and $\varepsilon_{i,t}$ the residuals.

Model 4 tests H1c. Here, we interact *Workshop* with *Player Numbers* to examine whether the effect of *Workshop* on price changes with more players or equivalently phrased, whether the effect of *Player Numbers* on price changes, given the existence of *Workshop* in a game.

$$\begin{aligned} Price_{i,t} = \beta_0 &+ \beta_1(Workshop_i * Age_{i,t}) \\ &+ \beta_2(Tradingcards_i * Age_{i,t}) \\ &+ \beta_3(Achievements_i * Age_{i,t}) \\ &+ \beta_4(Workshop_{i,t} * PlayerNumbers_{i,t}) \\ &+ \beta_5 Age_{i,t} + \beta_6 PlayerNumbers_{i,t} + u_i + \varepsilon_{i,t} \end{aligned}$$

where i denotes a specific game, t a specific day, u_i the individual (fixed) effect of a game, and $\varepsilon_{i,t}$ the residuals. To cater for potential heteroskedasticity and autocorrelation, we cluster Standard Errors on a game level in Model 3 and 4 (White, 1980, pp. 817). An alternative, which is not explored in the scope of this paper, is to use a model based on Baltagi (1999) who proposes a Fixed Effects Regression with Autoregression Type 1 Disturbance, suitable for large time dimensions.

6. Estimation Results

Table 2 and Table 3 at the end of this chapter provide a condensed overview of our results. Full reports are found in the Appendix. We present the results in order of our hypotheses.

H1a argued that games that include the possibility of User Generated Content via *Workshop* are priced at a higher level. We test this at two distinct timepoints: Model 1 looks at the price at release, while Model 2 looks at the price at present (March 2020). Our results show that games including *Workshop* demonstrate an on average 17% significantly higher release price and an on average 16% significantly higher present price, offering consistent support for H1a. We also found support for H1b in Model 3. Specifically, games with *Workshop* have on average 0.2% less price decline every month compared to games without (see Figure 7).

Lastly, in Model 4 we found a significant positive interaction between *Player Numbers* and *Workshop*, showing that

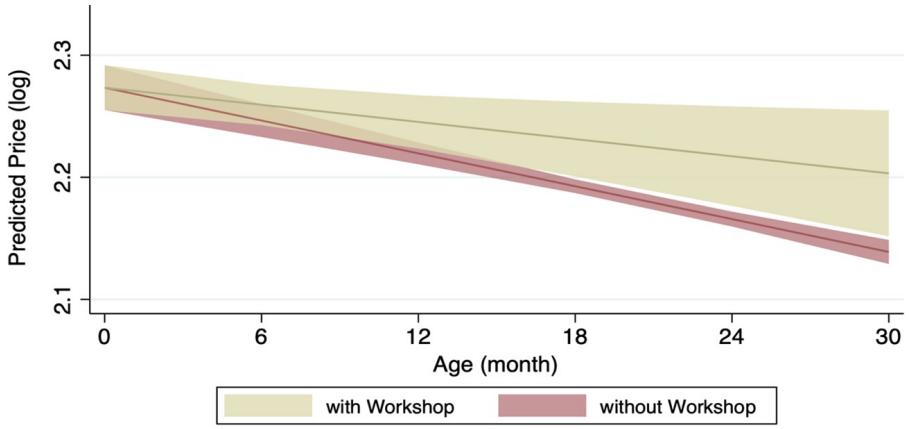


Figure 7: Price Development depending on Workshop Feature (95% CIs)

the effect of Player Number on price changes when the game includes Workshop, thus offering support for H1c (see Figure 9).

H2a argued that games that include the possibility of collecting and trading virtual items in Form of *Tradingcards* are priced at a higher level. Again, we tested this at release (Model 1) and present (Model 2). Contrary to our expectations, the feature *Tradingcards* exhibits a significantly negative effect on both the release price and present price. Games with *Tradingcards* have a on average 27% lower release price and 33% lower present price, leading us to reject H2a. Finally, we find with Model 3 that games with *Tradingcards* are subject to a greater price decline over time than games without (see Figure 10), which leads us to the reject H2b. Specifically, games with *Tradingcards* show an average of 0.3% more price decline every month compared to games without.

H3a argued that games that include the possibility of completing in-game challenges and earning social status via *Achievements* are priced at a higher level. We find a positive but, at the 5% level, just insignificant effect on release price (Model 1). However, the feature demonstrates a positive significant effect on present price, i.e. games with *Achievements* have a on average 12% higher present price (Model 2). This implies that H1c is only supported at an advanced life-cycle stage of a game. H3b is confirmed (Model 3), as we see less price decline with age for games with the feature (see Figure 11). Specifically, *Achievements* games show an average of 0.23% less price decline every month.

Beyond our Hypotheses, we also estimated the effect of *Player Numbers* on price and assumed that it changes over the life-cycle of a game, by interacting it with *Age*. We found that *Player Numbers* is negatively associated with price over most of a game's lifetime (Figure 13). In the Limitations (see Chapter 7), we will address this fact critically.

Finally, we conducted robustness checks on all models. For Model 1 and 2 by firstly taking up a reduced number of variables and secondly extending them with variables that represent less tangible user-defined characteristics about a game such as "Diffcult", "Classic" or "Family Friendly" (see Appendix Table A6, Table A7). For Model 3 and 4 we took up a reduced number of Variables (see Appendix Table A8). All checks show consistent results.

A condensed summary of previously discussed results is found in Table 2 and Table 3.

7. Discussion and Limitations

7.1. Discussion

In H1 we argued that developers who allow User Generated Content by including *Workshop* functionality can (a) charge higher prices compared to other games and (b) delay the natural price decline of their game. In essence, we justified this by the fact that players value the possibility to customize a game with their own and peers' content and thus keeping it engaging and stimulating in the long run. We wanted to know whether this higher perceived value can be exploited by developers in the form of higher prices. We tested H1a at two distinct time points: at release (Model 1), representing purely the developers perception of value, and at present (Model 2), representing a shared belief about the game, as players have been able to provide price feedback through either buying or not. We found consistent support for H1a by observing a higher price both at release (Model 1) and present (Model 2). Additionally, we found H1b, i.e. less price decline over time, confirmed, indicating that User Generated Content continually rejuvenates the game as it technically ages (see Figure 7). The support for H1c additionally shows that *Workshop* functionality provides users with network utility via indirect network effects. The effect of *Workshop* increases when a game has a larger player community, most likely as each player has more peer-content to integrate (see Figure 9).

In H2 we argued that developers who allow players to collect and trade virtual items in form of *Tradingcards* can (a) charge higher prices compared to other games and (b) delay the natural price decline of their game. Essentially, we hypothesized that players derive, amongst others, sentimental, monetary and social (e.g. status) value that developers exploit through higher prices. To our surprise, we found *Tradingcard* games have a lower price level, both at

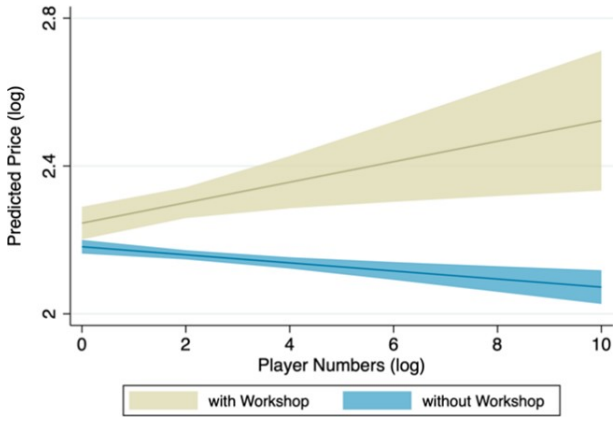


Figure 8: Price by Players dep. on Workshop

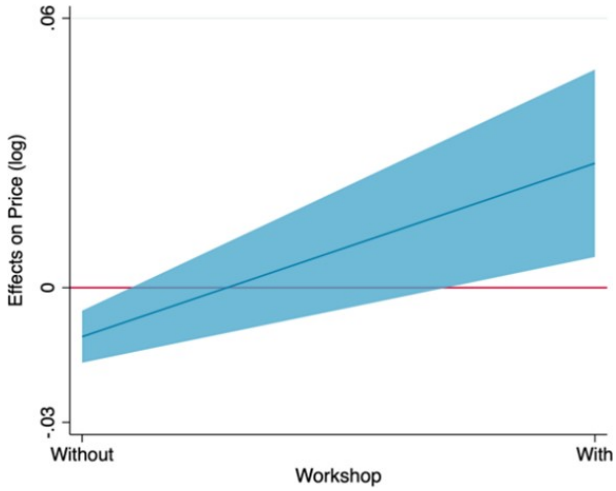


Figure 9: Effect of Players dep. on Workshop

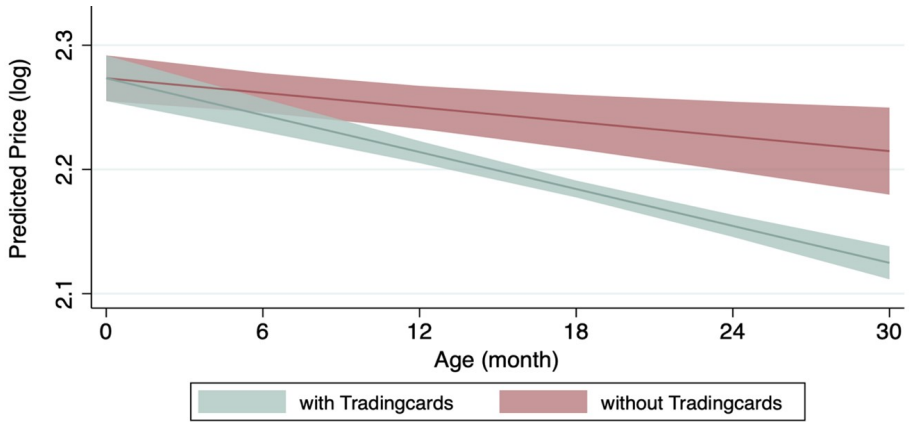


Figure 10: Price Development depending on Tradingcards Feature (95% CIs)

release and present. Additionally, they are also associated with a stronger price decline (see Figure 10). These finding are in sharp contrast to our theory and previous literature. Therefore, we venture another look at the mechanics of the feature to identify potential oddities that change the way the feature works in the context of Steam. We discovered that in the past, most recently in November 2019, Steam has banned large numbers of games that, as they explain vaguely, were "abusing Steam tools for financial gain" (PC Gamer, 2019). Through further research we found Reddit threads and an old Steam statement suggesting that some developers released low-cost games with *Tradingcards* functionality to distribute

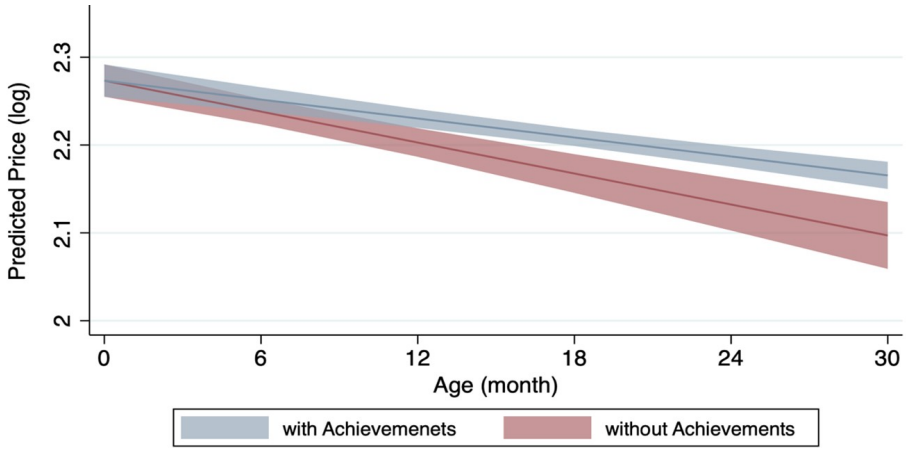


Figure 11: Price Development depending on Achievements Feature (95% CIs)

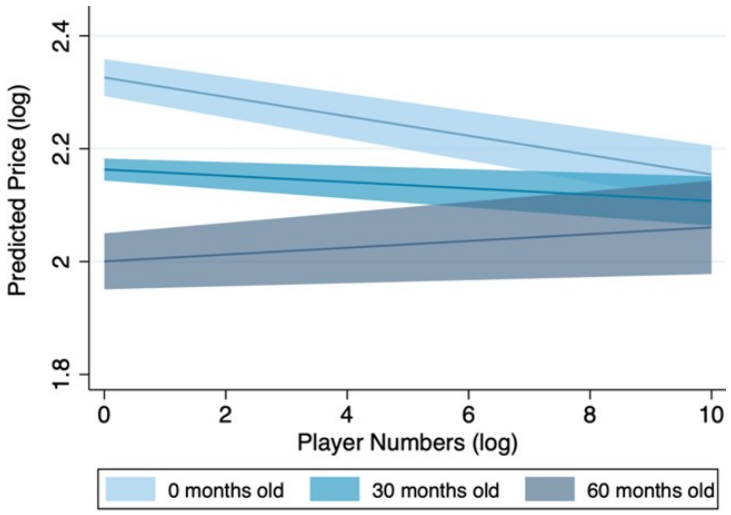


Figure 12: Price by Players for different Age

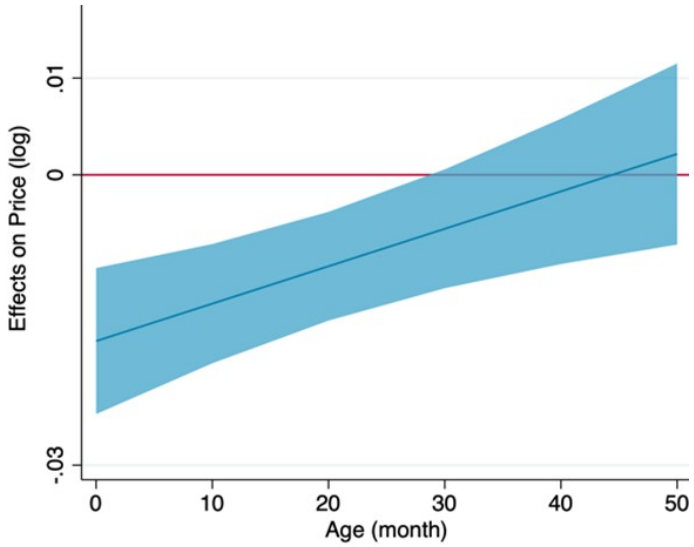


Figure 13: Effect of Players dep. on Age

Table 2: Effect of Community Features on Price Level

Model	1		2	
IndependentVariable	Price Release (log)		Price Present (log)	
Workshop	0.16**	(0.05)	0.15**	(0.06)
Tradingcards	-0.32***	(0.04)	-0.40***	(0.05)
Achievements	0.09	(0.04)	0.12*	(0.05)
N	2,029		1,763	
Adj.R^2	0.66		0.68	

Note: Full report see Appendix Table A6, Table A7. Robust SE in parentheses. ***p<0.001, **p<0.01, *p<0.5.

Table 3: Effect of Community Features on Price Development

Model	3		4	
IndependentVariable	Price (log)		Price (log)	
Age	-0.005***	(0.001)	-0.004***	(0.001)
*Age*Workshop*	0.002*	(0.001)	0.003**	(0.001)
*Age*Tradingcards*	-0.003***	(0.001)	-0.004***	(0.001)
*Age*Achievements*	0.002**	(0.001)	0.003**	(0.001)
PlayerNumbers	-0.017***	(0.004)	-0.011***	(0.003)
*PlayerNumbers*Age*	-0.000***	(0.000)		
*PlayerNumbers*Workshop*			0.039***	(0.011)
N	1,799,274		1,799,274	
Adj.R^2	0.08		0.07	

Note: Full report Appendix Table A8. Robust SE clustered on Game-level in parentheses. Age in Month, PlayerNumbers in log. ***p<0.001, **p<0.01, *p<0.5.

cards en masse to players or bots, and then profit from each trade through fees on their cards (Reddit, 2019; Steam Blog, 2017). A closer look at our sample confirms that in November a disproportionate amount or games that retired or, as we now know, were banned had *Tradingcards* as their only feature (see Figure 14). While this fact has compromised the expected effect of *Tradingcards* in our analysis, it does shed light on the importance of designing robust incentive systems when developing Community Features.

In H3, we argued that developers who allow players to complete in-game challenges and thereby earn social status via *Achievements* can (a) charge higher prices compared to other games and (b) delay the natural price decline of their game. At its core we justified this logic by assuming increased perceived value through serving intrinsic needs such as joy, feelings progression and accomplishment (see also goal-setting theory and reinforcement-theory) on the one hand, and extrinsic needs such as signaling competency and building social status on the other hand. In terms of H3a, we did not find significantly higher prices at the time of release, but at present. This discrepancy may indicate that developers, when setting the release price, initially underestimate the value that players gain from such functionality. A different explanation may be that developers also consider the amount of effort required to integrate certain functionalities. In contrast to *Workshop*, *Achievements* are very easy to integrate by defining a few simple challenges for the player. Lastly, we found H3b, i.e. less price decline over time, confirmed, indicating that challenges offer long-term value to the game (see Figure 11). This makes sense, since especially the extrinsic value (social status) that players derive from collecting *Achievements* is independent of game age, hence players are still interested in games at later life stages.

7.2. Limitations

First, in this study we investigate the influence of community features on pricing of games. It is based on the logic that ceteris paribus increased utility is expressed in increased willingness to pay, which developers capture through higher prices. The validity of this rationale is based on two assumptions. First, that the market is largely efficient, since only then will the equilibrium price reflect both supply and demand (aggregated willingness to pay). Although we show that digital gaming platforms share many elements of perfect markets (see Chapter 3: Industry Background), we are aware that no real-world market is perfect. Secondly, price is not only a measure of value but also a marketing tool. Thus, it is conceivable that developers may pursue penetration strategies via initial discounts or even pursue alternative monetization models. We control as much as possible

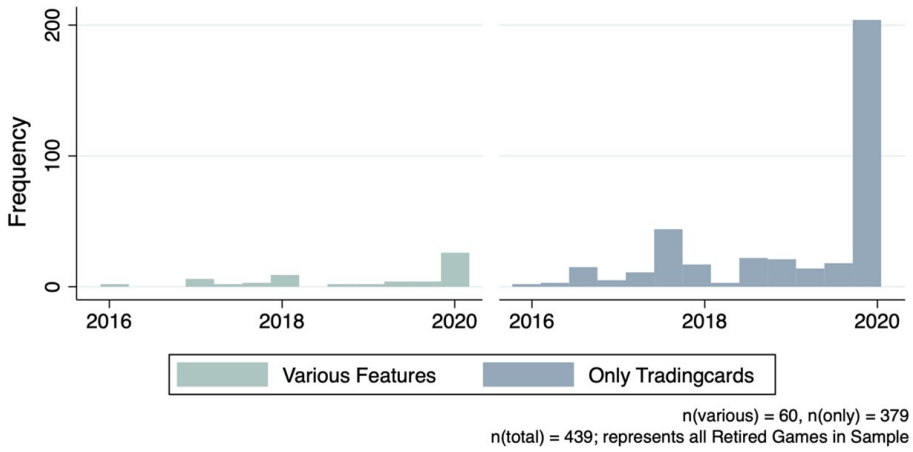


Figure 14: Distribution of Retired Games in Sample differentiated by Features

for such strategies by using, for example, the base price of a game, which ignores strategic short-term discounts and only reflects long-term price changes (see Figure 3). In addition, we control as much as possible for complementary monetization strategies, such as additional paid content (DLCs) or In-App-Purchases. Free games are completely excluded. Despite these measures, it is inevitable that unobserved strategies impact pricing.

Second, we looked at a very concrete setting, which limits the generalizability of our results. Community Involvement is a critical success factor for the longevity of products in the PC video game market. However, the success factor "Community" may be less important for other digital platforms, where content consumption involves less interaction with other consumers or producers, such as music and video streaming platforms. Furthermore, long content lifecycles are not of the same relevance for all digital platforms. For news platforms, for example, speed is key while fixed cost are comparably low. Additionally, we also have to be careful about generalizing among PC video games, as we restricted our analysis to the most popular games on Steam due to data quality.

Third, we initially tested the assumptions of Random Effects (see Chapter 5: Model Specification) to be able to estimate the effect of time-invariant game features more easily, however, had to proceed with Fixed Effects. We solved this problem by interacting such variables with time-variant variables to still be able to make statements about price development. Still, it would be worth testing for Random Effects with adapted model specifications (e.g. excluding/including variables) or look into Mixed Models (UCLA, 2020).

Forth, we found that the *Player Number* is negatively associated with price (Model 3). We acknowledge that price and *Player Numbers* may influence each other. When prices are low, for example when discounts are offered, more players are attracted. We identified two solutions to find out more about the causal relationship. First, we could include an instrument variable that has high correlation with *Player Numbers* and low correlation with price to perform a Two Stage OLS. The number of *Twitch Viewers* for each game over time could be suitable. Second, we could lag the variable *Player Numbers* by one month, which gives us an indication of how developers are monetizing a larger player base. In the scope of this paper, we tested Model 3 again without the variable *Player Numbers* to make sure the coefficients of the Community Features were not biased.

Lastly, we have to interpret the results of Model 1 and 2 conservatively, as we do not use Fixed Effects, that absorb any heterogeneity between games. Consequently, there might be unobserved differences between games, e.g. team size or budget, which we cannot account for despite many control variables (omitted variable bias). It would therefore be inappropriate to speak of a clear causal effect of Community Features on price, and more correctly to speak of associations. Additionally, the study assumes that theoretically every game could equally integrate Community Features. However, it is possible that Community Features make less sense for some games, perhaps in the case of story-driven single-player games. At the same time, Community Features may be particularly useful in the context of multiplayer games, which may already have an active, self-sustaining player base, making it inappropriate to attribute positive effects on price to features alone. In reality, therefore, integration is not always decided randomly, but depends on the nature of the game.

8. Conclusion, Implications and Future Research

8.1. Conclusion

In this paper, we set out to answer the question of how Community Involvement influences the price setting of firms on digital platforms. We addressed this question in the context of the gaming platform Steam by looking at three novel ways to promote player interaction and involvement. Specifically, we investigated how pricing of games is influenced by three Community Features that allow players to (1) customize a game with own and peers' game content via the *Workshop* Feature, (2) collect and trade virtual items via the *Tradingcards* Feature, (3) completing challenges and earning

social status via the *Achievements* Feature. We showed that games with *Workshop* or *Achievements* demonstrate a higher price level, the latter only at a more advanced age, as well as less price decline over time. Contrary to expectations, we found games with *Tradingcards* to be lower priced and to have stronger price decline, which we explained by the "misuse" of the feature by game developers. With regard to our research question, this predominantly suggest that game developers can charge higher prices for games with Community Features and secondly, that they can delay the natural price decline of their game.

8.2. Implications

Our findings are relevant for game developers. First, they should consider integrating Community Features as a means of maintaining consumer utility and thus willingness to pay in the long run. However, as discussed before (see Chapter 7: Limitations), Community Features do not fit all games - e.g. a multiplayer game may benefit from them differently than a single-player game. The assumption that the mere integration of these features allows to charge higher prices is misguided. Rather, the nature of the game must be taken into account. In the best case, a natural fit between game mechanics and Community Features is already ensured at the conception stage of the game. Secondly, developers who seek a two-way relationship and in particular co-creation with the consumer must be prepared for new time-consuming and costly tasks. They need to respond to customer requests, integrate feedback, provide development tools, create guidelines and ensure compatibility of User Generated Content. Therefore, it is important to consider the trade-off between potentially higher revenue and increased costs of implementing and maintaining Community Features.

Our findings are also important for platform owners. First, we know that the economic success of platforms and content producers is closely linked through revenue sharing. Platforms should therefore think about helping their complementors succeed by empowering them with powerful Community Features that increase the consumer's value. Second, platforms often fight for complementors and consumers according to the "Winner-Takes-It-All" principle. With attractive Community Features, platforms could differentiate themselves in competition with other platforms.

8.3. Future Research

Future research could help generalize our findings for the PC video gaming industry. For example, the sample should be increased and randomized to investigate if our findings apply to games that are less popular. It would also be interesting to investigate the effect of Community Features on the size of the player base of a game over time. In light of growing popularity of alternative monetization strategies, such as recurring revenue through Micro-Transactions, In-App-Purchases, Subscriptions, DLCs, sustaining a loyal player base is essential. Relatedly, it would also be worth looking into the rising phenomenon of free games and their strategies of attracting and monetizing on players. Lastly, it is worth further dissecting the effect of Community Features in inherent utility that's derived regardless of how many others use the features (e.g. customization through own User Generated Content) and network utility that's dependent on how many others use the feature as well (e.g. integration of peers' User Generated Content).

More importantly, we encourage future research of Community Involvement in the context of other novel digital platforms. It will be interesting to see if the new types of Community Involvement we investigated are equally effective in case of media products like music, video and news that have different life-cycles and communities and consumption behavior.

References

Adner, R., & Levinthal, D. (2001). Demand Heterogeneity and Technology Evolution: Implications for Product and Process Innovation. *Management Science*, *47*(5), 611–628.

Baltagi, B. H. (1999). Unequally spaced panel data regressions with AR (1) disturb-ances. *Econometric Theory*, *15*(6), 814–823.

Banks, J. (2013). *Co-creating Videogames*. London: Bloomsbury Publishing.

Barnett, J., & Coulson, M. (2010). Virtually Real: A Psychological Perspective on Massively Multiplayer Online Games. *Review of General Psychology*, *14*(2), 167–179.

Bass, F. M. (1980). The Relationship Between Diffusion Rates, Experience Curves, and Demand Elasticities for Consumer Durable Technological Innovations. *The Journal of Business*, *53*(3), 51–67.

Bass, F. M., & Bultez, A. V. (1982). Technical Note-A Note on Optimal Strategic Pric-ing of Technological Innovations. *Marketing Science*, *1*(4), 371–378.

Benlian, A., Vetter, J., & Hess, T. (2010). Pricing of Content Services-An Empirical Investigation of Music as a Service. In *16th americas conference on information sys-tems.*

Böhm, S., Adam, F., & Farrell, W. C. (2015). Impact of the mobile operating system on smartphone buying decisions: A conjoint-based empirical analysis. In *International conference on mobile web and information systems* (pp. 198–210). Springer Verlag.

Bork, R. H. (1993). *The Antitrust Paradox: A Policy at War with Itself.* New York: Free Press.

Bostan, B. (2009). Player motivations: A psychological perspective. *Computers in Entertainment*, *7*(2), 32.

Bowen Martin, C., & Deuze, M. (2009). The Independent Production of Culture: A Digital Games Case Study. *Games and Culture*, *4*(3), 276–295.

Breusch, T. S., & Pagan, A. R. (1979). A Simple Test for Heteroscedasticity and Ran-dom Coefficient Variation. *Econometrica*, *47*(5), 1287–1294.

Burger-Helmchen, T., & Cohendet, P. (2011). User Communities and Social Software in the Video Game Industry. *Long Range Planning*(5), 44.

Cadin, L., Guérin, F., & Defillippi, R. (2006). HRM Practices in the Video Game In-dustry:. Industry or Country Contingent? *European Management Journal*, *24*(4), 288–298.

Castronova, E. (2005). Synthetic Worlds: The Business and Culture of Online Games. Chicago: University of Chicago Press.

CERI. (2006). *Centre for Educational Research and Innovation.* Paris: OECD Publishing (Schooling for Tomorrow).

Chatterjee, R. (2009). Strategic pricing of new products and services. In *Handbook of pricing research in marketing* (pp. 169–215). Northhampton: Edward Elgar Publishing.

Chatterjee, S., & Hadi, A. S. (1986). Influential Observations, High Leverage Points, and Outliers in Linear Regression. *Statistical Science*, *1*(3), 379–393.

Chesbrough, H. W. (2003). *Open Innovation The New Imperative for Creating and Prof-iting from Technology.* Boston: Harvard Business School Press.

Clarefield. (2018). Gaming Industry: Facts, Figuers and Trends. *Journal of Social Behavior & Personality*.

Claussen, J., Kretschmer, T., & Spengler, T. (2010). *Backward compatibility to sustain market dominance-evidence from the US handheld video game industry.*

Csikszentmihalyi, M. (1990). *FLOW: The Psychology of Optimal Experience.* New York: Harper.

Danet, B., & Katriel, T. (1994). No two alike: play and aesthetics in collecting. In *-terpreting objects and collections. london: Routledge* (pp. 220–239).

Dean, J. (1976). Pricing Policies for New Products. *Harvard Business Review*.

Denny, P. (2013). The effect of virtual achievements on student engagement. In *Conference on Human Factors in Computing Systems - Proceedings* (pp. 763–772).

Dockner, E. J., & Fruchter, G. E. (2004). Dynamic strategic pricing and speed of diffusion. *Journal of Optimization Theory and Applications*, *123*(2), 331–348.

Dolan, R. J., & Jeuland, A. P. (1981). Experience Curves and Dynamic Demand Mod-els: Implications for Optimal Pricing Strategies. *Journal of Marketing. Winter*, *45*(1), 52–62.

Eisenmann, T., Parker, G., & Van Alstyne, M. W. (2006). Strategies for Two-Sided Markets. *Harvard Business Review*.

Farzan, R., & et al. (2008). Results from Deploying a Participation Incentive Mechanism within the Enterprise. In *Sigchi conference on human factors in computing systems.*

Formanek, R. (1994). Why they collect: Collectors reveal their Motivations. In *Inter-preting objects and collections* (pp. 327–335).

Franke, N., & Von Hippel, E. (2002). *Satisfying Heterogeneous User Needs via Inno-vation Toolkits: The Case of Apache Security Software.*

Fruchter, G., & Orbach, Y. (2008). A utility-based dynamic model used to predict ab-normalities in diffusion over time. *Innovative Marketing*, *4*(1), 37–45.

Gallagher, S., & Park, S. H. (2002). Innovation and Competition in Standard-Based Industries: A Historical Analysis of the U.S. Home Video Game Market. *IEEE Trans-actions on Engineering Management*, *49*(1), 67.

Golder, P. N., & Tellis, G. J. (2003). Growing, Growing, Gone: Cascades, Diffusion, and Turning Points in the Product Life Cycle. *Marketing Science*, *23*(2), 207–218.

Gundlach, H., & Hofmann, J. (2017). Preferences and willingness to pay for tablet news apps. *Journal of Media Business Studies*, *14*(4), 257–281.

Guo, Y., & Barnes, S. (2007). Why People Buy Virtual Items in Virtual Worlds with Real Money. In *Data base for advances in information systems* (Vol. 38, pp. 69–76).

Hahsler, M., & Breidert, C. (2006). Adaptive Conjoint Analysis for Pricing Music Downloads. In *Advances in Data Analysis, Proceedings of the 30th Annual Conference of the Gesellschaft für Klassifikation* (pp. 409–416).

Hallford, N., & Lamothe, A. (2001). *Swords and circuitry: A designer's guide to com-puter role-playing games.* Portland: Premier Press.

Hamari, J. (2013). Transforming homo economicus into homo ludens: A field experi-ment on gamification in a utilitarian peer-to-peer trading service. *Electronic Commerce Research and Applications*, *12*(4), 236–245.

Hamari, J., & Eranti, V. (2011). Framework for Designing and Evaluating Game Achievements. In *Digra 2011 conference: Think design play.*

Hamari, J., & Koivisto, J. (2013). Social motivations to use gamification: an empirical study of gamifying exercise. In *21st european conference on information systems.*

Hamari, J., & Lehdonvirta, V. (2010). Game design as marketing: How game mechan-ics create demand for virtual goods. *Journal of Business Science and Applied Management*, *5*(1), 14–29.

Hausman, J. A. (1978). Specification Tests in Econometrics. *Econometrica*, *46*(6), 1251–1271.

Humphreys, S. (2005). Productive users, intellectual property and governance: the chal-lenges of computer games. *Media and Arts Law Review*, *10*(4), 299–310.

Jakobsson, M. (2011). The Achievement Machine: Understanding Xbox 360 Achieve-ments in Gaming Practices. *The International Journal of Computer Game Research*, *11*(1), 1–22.

Jakobsson, M. (2016). *'Achievements', in Debugging Game History: A Critical Lexicon.* Boston: MIT Press.

Jeppesen, L. B. (2004). *Profiting from innovative user communities: How firms organize the production of user modifications in the computer games industry.*

Jeppesen, L. B. (2005). User Toolkits for Innovation: Consumers Support Each Other. *Journal of Product Innovation Management*, *22*(4), 347–362.

Jöckel, S., Will, A., & Schwarzer, F. (2008). Participatory Media Culture and Digital Online Distribution-Reconfiguring the Value Chain in the Computer Game Industry. *International Journal on Media Management*, *10*(3), 102–111.

Jung, J. H., Schneider, C., & Valacich, J. (2010). Enhancing the Motivational Af-fordance of Information Systems: The Effects of Real-Time Performance Feedback and Goal Setting in Group Collaboration Environments. *Management Science*, *56*(4), 724–742.

Juul, J. (2010). *A casual revolution: Reinventing video games and their players.* Boston: MIT press.

Kalish, S. (1983). Monopolist Pricing with Dynamic Demand and Production Cost. *Source: Marketing Science*, *2*(2), 135–159.

Katz, M. L., & Shapiro, C. (1985). Network Externalities, Competition, and Compatibility. *The American Economic Review*, *75*(3), 424–440.

Koch, S., & Bierbamer, M. (2016). Opening your Product: Impact of User Innovations and their Distribution Platform on Video Game Success. *Electronic Markets*, *26*(4), 357–368.

Koch, S., & Kerschbaum, M. (2014). Joining a Smartphone Ecosystem: Application Developers' Motivations and Decision Criteria. *Information and Software Technology*, *56*(11), 1423–1435.

Krishnan, T. V., Bass, F. M., & Jain, D. (1999). Optimal Pricing Strategy for New Products. *Management Science*, *45*(12), 1650–1663.

Landsman, V. and Stremersch, S. (2011). Multihoming in Two-Sided Markets: An Empirical Inquiry in the Video Game Console Industry. *Journal of Marketing*, *75*(6), 39–54.

Lastowka, G. (2007). User-Generated Content and Virtual Worlds. *Vanderbilt Journal of Entertainment and Technology Law*, *10*(1), 893–912.

Leibovitz, L. (2013). Atari Bankrupt: Video Game Industry Refuses to Heed History. *The New Republic*.

Lerner, J., & Triole, J. (2002). The Simple Economics of Open Source. *Journal of Industrial Economics*, *52*(1), 197–234.

Levitt, T. (1965). Exploit the Product Life Cycle. *Harvard Business Review*.

Lilien, G., & et al. (2001). Performance Assessment of the Lead User Idea Generation Pro-cess for New Product Development. *Management Science*, *48*(8), 1054–1059.

Lin, D., Bezemer, C. P., & Hassan, A. E. (2018). An empirical study of early access games on the Steam platform. *Empirical Software Engineering*, *23*(2), 771–799.

Liu, H. (2010). Dynamics of Pricing in the Video Game Console Market: Skimming or Penetration? *Journal of Marketing Research*, *47*(3), 428–443.

Locke, E. A., & Latham, G. P. (1991). A Theory of Goal Setting and Task Performance. *The Academy of Management Review. The Academy of Management*, *16*(2), 212–247.

Locke, E. A., & Latham, G. P. (2002). Building a Practically Useful Theory of Goal Setting and Task Motivation A 35-Year Odyssey. *American Psychologist. American Psychological Association Inc*, *57*(9), 705–717.

Luthans, F., & Stajkovic, A. D. (1999). Reinforce for Performance: The Need to go Beyond Pay and even Rewards. *The Academy of Management Executive*, *13*(2), 49–57.

Marchand, A. (2016). The power of an installed base to combat lifecycle decline: The case of video games. *International Journal of Research in Marketing*, *33*(1), 140–154.

Mekler, E. D., Brühlmann, F., Opwis, K., & Tuch, A. (2013). *Do Points, Levels and Leaderboards Harm Intrinsic Motiva-tion? An Empirical Analysis of Common Gamification Elements*. Ontario.

Meyer, T. G. (2012). *Path Dependence in Two-Sided Markets*.

Montola, M., Nummenmaa, T., Lucero, A., Boberg, M., & Korhonen, H. (2009). Applying Game Achievement Systems to Enhance User Experience in a Photo Sharing Service. In *13th international academic mindtrek conference: Everyday life in the ubiquitous era* (pp. 94–97).

Nair, H. (2004). Intertemporal Price Discrimination with Forward-Looking Consumers: Application to the US Market for Console Video-Games. *Quantitative Marketing and Economics*, *5*(3), 239–292.

Nikou, S., & et al. (2014). A Consumer Perspective on Mobile Service Platforms: A Con-joint Analysis Approach. *Communications of the Association for Information Systems*, *34*(1), 1409–1424.

Osborne, J. (2010). Improving your data transformations. *Practical Assessment*, *15*(1), 12.

Park, B. W., & Lee, K. C. (2011). Exploring the value of purchasing online game items. *Computers in Human Behavior*, *27*(6), 2178–2185.

Park, M. (2010). Defying Obsolescence. In *Longer lasting products: Alternatives to the throwaway society* (p. 77). Vermont: Gower Publishing.

Parmentier, G., & Gandia, R. (2013). Managing Sustainable Innovation with a User Community Toolkit: The Case of the Video Game Trackmania. *Creativity and Innova-tion Management*, *22*(2), 195–208.

PC Gamer. (2019). *Valve removed 1000 games from Steam because publishers were 'abusing' Steamworks.* Retrieved 2020-06-14, from https://www.pcgamer.com/uk/1000-games-were-just-removed-from-steam-and-we-dont-know-why/

Poor, N. (2014). Computer game modders' motivations and sense of community: A mixed-methods approach. *New Media & Society*, *16*(8), 1249–1267.

PwC. (2017). *Global Entertainment and Media Outlook 2017.* Retrieved from www.pwc.com/outlook

PwC. (2019). *Global Entertainment and Media Outlook 2019.* Retrieved from https://www.pwc.com/gx/en/entertainment-media/outlook-2019/entertainment-and-media-outlook-perspectives-2019-2023.pdf

Reddit. (2019). *Steam is removing hundreds of games from the store.* Retrieved 2020-06-14, from https://www.reddit.com/r/Steam/comments/e1obe3/steam_is_removing_hundreds_of_games_from_the/

Rietveld, J., & Eggers, J. P. (2018). Demand Heterogeneity in Platform Markets: Implications for Complementors. *Organization Science*, *29*(2), 304–322.

Roberts, J. H., & Lilien, G. L. (1993). Explanatory and Predictive Models of Consumer Behavior. *Handbooks in Operations Research and Management Science*, *5*(1).

Rozin, P. (2000). Preadaptation and the puzzles and properties of pleasure. *Well-being: The foundations of hedonic psychology*, 109–133.

Scharkow, M. e. a. (2015). Beyond the "core-gamer": Genre preferences and gratifications in computer games. *Computers in Human Behavior*, *44*(1), 293–298.

Schreier, M. e. a. (2007). Lead users and the adoption and diffusion of new products: Insights from two extreme sports communities. *Marketing Letters*, *18*(2), 15–30.

Shy, O. (2001). The Economics of Network Industries. *Cambridge University Press*.

Shy, O. (2011). A Short Survey of Network Economics. *Review of Industrial Organization*, *38*(2), 119–149.

Steam. (2020a). *Features (Steamworks Documentation)*. Retrieved 2020-06-08, from https://partner.steamgames.com/doc/features

Steam. (2020b). *Steam Community: Steam Trading Cards.* Retrieved from https://steamcommunity.com/tradingcards/

Steam. (2020c). *Steam Direct Product Submission Fee on Steam.* Retrieved 2020-06-09, from https://store.steampowered.com/sub/163632

Steam. (2020d). *Steam Subscriber Agreement.* Retrieved 2020-06-21, from https://store.steampowered.com/subscriber_agreement/?l=english

Steam Blog. (2015). *Steam Blog: Mods on Steam.* Retrieved 2020-06-21, from https://store.steampowered.com/about/communitymods/?l=english

Steam Blog. (2017). *Changes to Trading Cards.* Retrieved 2020-06-14, from https://steamcommunity.com/games/593110/announcements/detail/1954971077935370845

Steam Blog. (2020). *Steam Blog: Removing Payment Feature Skyrim Workshop.* Retrieved 2020-06-21, from https://steamcommunity.com/games/SteamWorkshop/announcements/detail/208632365253244218

Steam Card Exchange. (2020). *Steam Card Exchange: Games with Tradingcards.* Retrieved 2020-06-22, from https://www.steamcardexchange.net/index.php?showcase

Steamspy. (2020). *Monthly summaries - SteamSpy - All the data and stats about Steam games.* Retrieved 2020-06-07, from https://steamspy.com/year/

Tabachnick, B. G., & Fidell, L. S. (2007). Using Multivariate Statistics (5th Edition). *Pearson Education*.

Thom, J., Millen, D., & DiMicco, J. (2012). Removing gamification from an enterprise SNS. *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, 1067–1070.

UCLA. (2020). *Introduction to Linear Mixed Models.* Retrieved 2020-06-24, from https://stats.idre.ucla.edu/other/mult-pkg/introduction-to-linear-mixed-models

Urban, G. L., & Von Hippel, E. (1988). Lead user analyses for the development of new industrial products. *Management Science*, *34*(5), 569–582.

Varian, H. (1992). *Microeconomic Analysis*. WW Norton & Company.

Von Hippel, E. (1986). Lead Users: A Source of Novel Product Consepts. *Management Science*, *32*(7), 791–805.

Von Hippel, E., & Katz, R. (2002). *Shifting Innovation To Users Via Toolkits* (Vol. 4232).

Wang, H., & Sun, C. T. (2011). Game Reward Systems: Gaming Experiences and Social Meanings. In *Digra 2011 conference.* Atlanta.

Weerahandi, S., & Dalal, S. R. (1992). A Choice-Based Approach to the Dif-

fusion of a Service: Forecasting Fax Penetration by Market Segments. *Marketing Science. Winter*, *11*(1), 39–53.

Whang, L. S. M., & Kim, J. Y. (2005). The Comparison of Online Game Experiences by Players in Games of Lineage & EverQuest: Role play vs. Consumption. In *DiGRA 2005 Conference: Changing Views - Worlds in Play*. Vancouver.

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, *48*(4), 817–838.

Yee, N. (2006). Motivations for Play in Online Games. *CYBERPSYCHOLOGY & BEHAVIOR*, *9*(6), 772–777.

Yusuf Dauda, S., & Lee, J. (2015). Technology adoption: A conjoint analysis of con-sumers' preference on future online banking services. *Information Systems*, *53*(1), 1–15.

Junior Management Science 6(3) (2021) 568-589

Junior Management Science

journal homepage: www.jums.academy

Cross-border Tax Group

Grenzüberschreitende ertragsteuerliche Organschaft

Leonie Rinke

Heinrich-Heine-Universität Düsseldorf

Abstract

The tax group for income tax purposes enables the income of a subsidiary to be attributed to the income of a parent company for corporate income tax and trade tax purposes. This makes it an important institution in German tax law, but in the cross-border case it harbors many ambiguities and problems. In this paper, therefore, the potential case constellations of a cross-border tax group are identified and examined for the possibility of forming a tax group. At the beginning, a brief description of the prerequisites and legal consequences of the tax group is given. Subsequently, a distinction is made between an inbound and an outbound situation as well as between a foreign domicile in an EU or EEA state and a foreign domicile in a non-member state. In particular, the problem areas in which the domestic application of law conflicts with EU case law are adressed. The results show that the ECJ case law has so far only been implemented by German courts to a limited extent. This leads to a legal situation which is to be classified as contrary to EU law, but which will probably only change through proceedings before the ECJ.

Zusammenfassung

Die ertragsteuerliche Organschaft ermöglicht in der Körperschaft- und Gewerbesteuer die Zurechnung des Einkommens eines Tochterunternehmens, der Organgesellschaft, zum Einkommen eines Mutterunternehmens, des Organträgers. Damit ist sie ein wichtiges Institut im deutschen Steuerrecht, birgt aber im grenzüberschreitenden Fall viele Unklarheiten und Probleme. In der vorliegenden Arbeit werden daher die potenziellen Fallkonstellationen einer grenzüberschreitenden Organschaft identifiziert und auf die Möglichkeit der Bildung einer Organschaft untersucht. Zu Beginn erfolgt eine kurze Darstellung der Voraussetzungen und Rechtsfolgen der Organschaft. Anschließend wird zwischen einem Inbound- und einem Outbound-Sachverhalt sowie zwischen einer Auslandsansässigkeit in einem EU- bzw. EWR-Staat und einer Auslandsansässigkeit in einem Drittland unterschieden. Es wird insbesondere auf die Problemfelder eingegangen, in welchen die inländische Rechtsanwendung mit EU-Rechtsprechung kollidiert. Im Ergebnis zeigt sich, dass die EuGH-Rechtsprechung von deutschen Gerichten bisher nur eingeschränkt umgesetzt wird. Dies führt zu einer Rechtslage, welche zwar als unionsrechtswidrig einzuordnen ist, welche sich jedoch wohl erst durch ein Verfahren vor dem EuGH ändern wird.[1]

Keywords: Organschaft; § 14 KStG; internationale Gruppenbesteuerung; Niederlassungsfreiheit.

1. Einleitung

Die vorliegende Arbeit beschäftigt sich mit den Möglichkeiten und Grenzen zur Bildung einer grenzüberschreitenden Organschaft. In einer internationaler werdenden Wirtschaftswelt werden Ländergrenzen für die Entscheidung über Investitionen immer unwichtiger, was dazu führt, dass die grenzüberschreitende Organschaft an Bedeutung gewinnt.[2] Problematisch ist in diesem Kontext, dass die Organschaft in Deutschland zwar schon eine lange Tradition hat, die betref-

[1] Das Rechtsprechungsverzeichnis, das Verzeichnis der Verwaltungsanweisungen, das Rechtsquellenverzeichnis sowie das Verzeichnis der sonstigen Quellen sind im Appendix dieser Arbeit aufgeführt. Dieser ist online unter www.jums.academy/l-rinke verfügbar.

[2] Vgl. Grotherr, Unzulänglichkeiten, StuW 1995, 125.

DOI: https://doi.org/10.5282/jums/v6i3pp568-589

fenden Normen jedoch im grenzüberschreitenden Fall häufig nicht präzise formuliert sind oder es zu Kollisionen mit Unionsrecht kommen kann.[3] Dies führt dazu, dass es bei grenzüberschreitenden Sachverhalten zu Streitigkeiten und Problemfällen bei der Umsetzung kommen kann, welche in der vorliegenden Arbeit genauer beleuchtet werden sollen.

Für die einzelnen Konstellationen wird im Folgenden jeweils zwischen einer Ansässigkeit in einem Mitgliedstaat der EU oder einem Vertragsstaat des EWR sowie einem Drittland unterschieden werden. Diese Gliederung ist darauf zurückzuführen, dass die der Organschaft zugrunde liegenden Normen diese Unterscheidung ebenfalls treffen.[4] EU-Mitgliedstaaten und Vertragsstaaten des EWR-Abkommens können gemeinsam behandelt werden, da sie im Gesetz die gleiche Würdigung erhalten und sich auf die gleichen Grundfreiheiten berufen können: EuGH-Urteile zur Gruppenbesteuerung in den Mitgliedstaaten befassten sich stets mit der Vereinbarkeit des nationalen Rechts mit der Niederlassungsfreiheit.[5] Die Niederlassungsfreiheit im EWR-Abkommen entspricht weitestgehend der Niederlassungsfreiheit im AEUV,[6] zudem ist für das Hoheitsgebiet der EU der EuGH für die Wahrung und Durchsetzung des EWR-Rechts zuständig und gem. Art. 3 Abs. 2 üGA auch für die sonstige Rechtsprechung im EWR-Raum zu berücksichtigen, was in der Praxis zu einer mit der Rechtsprechung des EuGH konformen Interpretation des EWR-Abkommens führt.[7] EFTA-Staaten können sich also inhaltlich aufgrund des EWR-Abkommens ebenso wie EU-Mitgliedstaaten auf die Grundfreiheiten des AEUV berufen.[8]

Das Rechtsinstitut der Organschaft ist das steuerrechtliche Pendant zum bilanzrechtlichen Konzern und besteht in der Körperschaftsteuer, der Gewerbesteuer, der Umsatzsteuer und der Grunderwerbsteuer.[9] In Deutschland existiert kein einheitliches Konzernsteuerrecht in dem Sinne, dass eine bilanzielle Konsolidierung unmittelbar zu einer steuerrechtlichen Konsolidierung führt, sondern es müssen eigene steuerrechtliche Voraussetzungen erfüllt sein, um letztere zu erreichen.[10] Der Begriff „Organschaft“ beschreibt einen bestimmten wirtschaftlichen Abhängigkeitstatbestand mit rechtlicher und faktischer Unterordnung eines Unternehmens unter ein anderes, wobei beide Unternehmen ihre zivil- und steuerrechtliche Selbständigkeit behalten.[11] Im Ergebnis soll der gemeinsamen Leistungsfähigkeit eines Konzerns dadurch Rechnung getragen werden, dass die Ergebnisse wirtschaftlich zusammengehöriger Unternehmen steuerlich zusammengefasst werden, indem das Einkommen der Tochtergesellschaft dem Gesellschafter zugerechnet und bei diesem besteuert wird.[12]

2. Grundlagen der Organschaft

2.1. Voraussetzungen

2.1.1. Persönliche Voraussetzungen

Um die Rechtsfolgen der Organschaft auszulösen, müssen bestimmte Voraussetzungen vorliegen. Diese kann man in persönliche Voraussetzungen, d.h. solche, die seitens der Organgesellschaft oder des Organträgers vorliegen müssen, und sachliche Voraussetzungen, d.h. solche, die zwischen den beteiligten Gesellschaften vorliegen müssen, unterteilen. Die Voraussetzungen für körperschaft- und gewerbesteuerliche Organschaft sind seit 2002 identisch, d.h. es liegt gem. § 2 Abs. 2 S. 2 GewStG genau dann eine gewerbesteuerliche Organschaft vor, wenn eine körperschaftsteuerliche Organschaft nach § 14 KStG vorliegt.[13]

Für die Organgesellschaft kommen gem. § 14 Abs. 1 S. 1 KStG zunächst nur eine Europäische Gesellschaft (SE), eine AG oder eine KGaA als Gesellschaftsformen infrage, wenn diese Gesellschaften den Ort ihrer GL nach § 10 AO im Inland und ihren Sitz nach § 11 AO in einem Mitgliedstaat der EU oder einem Vertragsstaat des EWR haben.[14] § 17 Abs. 1 S. 1 KStG erweitert diesen Kreis jedoch auf andere Kapitalgesellschaften mit Ort der GL im Inland und Sitz in einem Mitgliedstaat der EU oder einem Vertragsstaat des EWR, was auch eine GmbH oder eine nach ausländischem Recht gegründete Kapitalgesellschaft einschließt.[15] Diese Beschränkung auf Kapitalgesellschaften bedeutet, dass nicht alle unbeschränkt körperschaftsteuerpflichtigen Personen die Möglichkeit haben, als Organgesellschaft Teil einer Organschaft zu werden, so fehlt etwa Genossenschaften oder Vereinen diese Option.[16]

[3] Vgl. van Lishaut, Reform, FR 2009, 1031; Frotscher, Organschaft, IStR 2011, 697; zur Rechtsentwicklung vgl. Neumann, in: Gosch, KStG, § 14 KStG, Rn. 10-12 (2015); Krumm, in: Blümich, KStG, § 14 KStG, Rn. 9-19 (Februar 2019); Walter, in: Bott/Walter, KStG, § 14 KStG, Rz. 42-54.1 (Februar 2020).

[4] Vgl. § 14 Abs. 1 S. 1 KStG; § 17 Abs. 1 S. 1 KStG.

[5] Vgl. etwa EuGH v. 31.02.2005, C-446/03, Rs. „Marks & Spencer“, EU:C:2005:763, Rn. 28; EuGH v. 06.09.2012, C-18/11, Rs. „Philips Electronics UK“, EU:C:2012:532, Rn. 16; EuGH v. 2.6.2014, C-39/13, C-40/13, C-41/13, Rs. „SCA Group Holding u.a.“, EU:C:2014:1758; Rn. 18.

[6] Vgl. Art. 49, 54 AEUV; Art. 31 Abs. 1, Art. 34 EWR-Abkommen; EuGH v. 03.02.2015, C-172/13, Rs. „Kommission/Vereinigtes Königreich“, EU:C:2015:50, Rn. 39; Holterhus, in: Krenzler/Herrmann/Niestedt, EU-Außenwirtschaftsrecht, 132a. EWR, Rn. 281, 283, 285 (Oktober 2019).

[7] Vgl. Art. 3 Abs. 2 üGA; Holterhus, in: Krenzler/Herrmann/Niestedt, EU-Außenwirtschaftsrecht, 132a. EWR, Rn. 99 f. u. 211 (Oktober 2019).

[8] Vgl. Meilicke/Sedemund, Manninen-Urteil, DB 2005, 2042; Scheunemann, Verlustberücksichtigung, IStR 2006, 152.; Sedemund/Sterner, Folgen, DStZ 2006, 32.

[9] Vgl. Walter, in: Bott/Walter, KStG, § 14 KStG, Rz. 1 (Februar 2020); Müller/Detmering/Lieber, Organschaft, 11. Aufl. 2020, VII f.

[10] Vgl. Kolbe, in: H/H/R, KStG, § 14 KStG, Anm. 10 (September 2016); Walter, in: Bott/Walter, KStG, § 14 KStG, Rz. 1 (Februar 2020).

[11] Vgl. BFH v. 01.08.1984, I R 99/80, BStBl. II 1985, 18, Rz. 9; BFH v. 07.12.1994, I K 1/93, BStBl. II 1995, 175, Rz. 33; Müller, in: M/S/O, KStG, §§ 14-19 KStG, Rz. 2 (2019); Krumm, in: Blümich, KStG, § 14 KStG, Rn. 1 (Februar 2019).

[12] Vgl. Kolbe, in: H/H/R, KStG, § 14 KStG, Anm. 1 (September 2016); Dölker, in: Dölker/Ergenziger/Feldgen u.a., KStG, § 14 KStG, Rz. 1 (15.12.2018).

[13] Vgl. § 2 Abs. 2 S. 2 GewStG; UntStFG v. 20.12.2001, BGBl. I 2001, 3858; Dölker, in: Dölker/Ergenziger/Feldgen u.a., KStG, § 14 KStG, Rz. 3 (02.01.2020).

[14] Vgl. § 14 Abs. 1 S. 1 KStG.

[15] Vgl. § 17 Abs. 1 S. 1 KStG; Müller, in: M/S/O, KStG, § 14 KStG, Rz. 31 (2019).

[16] Vgl. Müller, in: M/S/O, KStG, § 14 KStG, Rz. 32 (2019); Dörr, Organschaft, IStR 2004, 268.

Die Organgesellschaft muss im Gegensatz zum Organträger nicht gewerblich tätig sein.[17]

Der Organträger muss gem. § 14 Abs. 1 S. 1 Nr. 2 KStG eine natürliche Person oder eine nicht steuerbefreite Körperschaft, Personenvereinigung oder Vermögensmasse i.S.d. § 1 KStG oder eine originär gewerbliche PersG i.S.d. § 15 Abs. 1 S. 1 Nr. 2 EStG sein.[18] Dementsprechend reicht eine gewerblich geprägte PersG nach § 15 Abs. 3 Nr. 2 EStG explizit nicht aus.[19] Die gewerbliche Tätigkeit darf nach Meinung der Finanzverwaltung nicht nur geringfügig sein, woran es allerdings Kritik in der Fachliteratur gibt.[20] Das Erfordernis der Gewerblichkeit soll sicherstellen, dass das gesamte Ergebnis des Organkreises der Gewerbesteuer unterliegt, um zu vermeiden, dass sich einzelne Organgesellschaften durch Bildung einer Organschaft der Gewerbesteuer entziehen können.[21]

Die Beteiligung an der Organgesellschaft muss einer inländischen Betriebsstätte des Organträgers zugeordnet sein, welche sowohl nach innerstaatlichem Steuerrecht als auch nach DBA der inländischen Besteuerung unterliegt.[22] Faktisch wird hierdurch sichergestellt, dass die inländische Betriebsstätte auch nach dem jeweils geltenden DBA als eine solche gilt.[23] Die Kriterien für die Zuordnung zu der Betriebsstätte sind nicht klar definiert, es ist allerdings nach h.M. ein funktionaler Zusammenhang zur Geschäftstätigkeit der Betriebsstätte relevant.[24] Diese Zuordnung muss ununterbrochen während der gesamten Dauer der Organschaft vorliegen, was nach Meinung der Finanzverwaltung bedeutet, dass dies während der gesamten Mindestlaufzeit des GAV von fünf Jahren sowie anschließend während des gesamten Wirtschaftsjahres der Organgesellschaft der Fall sein muss.[25] Auch eine mehrstufige Beteiligung ist möglich; in diesem Fall muss die Beteiligung an der vermittelnden Gesellschaft der inländischen Betriebsstätte zuzuordnen sein.[26] Ist der Organträger eine PersG und mittelbar über weitere Personengesellschaften beteiligt, muss nach Literaturmeinung auf allen Stufen eine gewerbliche Tätigkeit nach § 15 Abs. 1 Nr. 1 EStG vorliegen.[27] Weiterhin muss die Beteiligung im Gesamthandsvermögen der PersG als Organträger liegen.[28]

2.1.2. Sachliche Voraussetzungen

Gem. § 14 Abs. 1 S. 1 Nr. 1 KStG muss der Organträger von Beginn des Wirtschaftsjahres der Organgesellschaft an ununterbrochen die Mehrheit der Stimmrechte an der Organgesellschaft halten.[29] Diese Voraussetzung, die sog. finanzielle Eingliederung, gilt auch für mittelbare Beteiligungen, diese sind aber nur dann zu berücksichtigen, wenn an jeder vermittelnden Gesellschaft die Mehrheit der Stimmrechte gehalten wird.[30] Ist die Stimmrechtsmehrheit auf jeder Beteiligungsebene gegeben, wird trotz Literaturkritik nicht die Additionsmethode angewandt, wodurch die volle Beteiligung dem Mutterunternehmen zugerechnet werden würde, da aufgrund der Mehrheit die Kontrolle gegeben ist.[31] Nach Meinung der Finanzverwaltung soll die Durchrechnungsmethode Anwendung finden, d.h. die Höhe der Beteiligungen auf jeder Ebene wird miteinander multipliziert.[32]

Eine weitere wichtige Voraussetzung für die Errichtung einer Organschaft ist der Abschluss und die Durchführung eines GAV i.S.d. § 291 Abs. 1 AktG über fünf Jahre hinweg.[33] Der GAV ist in §§ 291-307 AktG zunächst nur für die Kapitalgesellschaftsformen SE, AG und KGaA geregelt.[34] Für Kapitalgesellschaften anderer Rechtsformen sind §§ 14-16 KStG dennoch entsprechend anzuwenden, wenn diese sich zivilrechtlich wirksam verpflichten, ihren ganzen Gewinn an ein anderes Unternehmen abzuführen, diese Gewinnabführung den Höchstbetrag nach § 301 AktG nicht überschreitet und eine Verlustübernahme durch Verweis auf § 302 AktG vereinbart wird.[35] Die weiteren Voraussetzungen der Organschaft werden von dieser Sonderregelung nicht berührt.[36] Bei mehrstufigen Beteiligungen kann der GAV auch direkt zwischen Mutter- und Enkelgesellschaft unter Umgehung der zwischengeschalteten Gesellschaft(-en) abgeschlossen wer-

[17] Vgl. BFH v. 08.12.1971, I R 3/69, BStBl. II 1972, 289, Rz. 9.

[18] Vgl. § 14 Abs. 1 S. 1 Nr. 2 KStG; Dölker, in: Dölker/Ergenziger/Feldgen u.a., KStG, § 14 KStG, Rz. 15 (20.03.2020).

[19] Vgl. BMF v. 10.11.2005, IV B 7 - S 2770 - 24/05, BStBl. I 2005, 1038, Rn. 15; Blumers/Goerg, Gewerblichkeit, BB 2003, 2203; Schulze zur Wiesche, Organschaft, DStZ 2013, 622.

[20] Vgl. BMF v. 10.11.2005, IV B 7 - S 2770 - 24/05, BStBl. I 2005, 1038, Rn. 17; Förster, änderungen, DB 2003, 903; Füger, änderungen, BB 2003, 1758; Walter, Entwurf, GmbHR 2005, 458.

[21] Vgl. BFH v. 12.08.1965, IV 322/64 U; BStBl. III 1965, 589, Rz. 5 f.

[22] Vgl. § 14 Abs. 1 Nr. 2 S. 4-7 KStG.

[23] Vgl. Dötsch/Pung, in: D/P/M, KSt, § 14 KStG, Tz. 209 (Juni 2019); Dölker, in: Dölker/Ergenziger/Feldgen u.a., KStG, § 14 KStG, Rz. 22 (17.02.2020); Herbener, Holding, in: Prinz/Witt, Organschaft, 2. Aufl. 2019, 676, Rz. 17.9.

[24] Vgl. BFH v. 26.02.1992, I R 85/91, BStBl. II 1992, 937, Rz. 21; BMF v. 24.12.1999, IV B 4 - S 1300 - 111/99, BStBl. I 1999, 1076, Tz. 2.4; OFD Karlsruhe v. 16.01.2014, S 2770/52/2 - St 221, FR 2014, 434 zu § 14 Abs. 1 S. 1 Nr. 2 S. 6 und 7; Dölker, in: Dölker/Ergenziger/Feldgen u.a., KStG, § 14 KStG, Rz. 22 (07.02.2020); Prinz/Ludwig, Gewinnabführungsvertrag, DB 2020, 1023; genauer zur Zuordnung s.u. unter Punkt 3.1.1.2.

[25] Vgl. § 14 Abs. 1 S. 1 Nr. 2 S. 4, Nr. 3 S. 1 KStG; OFD Karlsruhe v. 16.01.2014, S 2770/52/2 - St 221, FR 2014, 434 zu § 14 Abs. 1 S. 1 Nr. 2 S. 4 und 5.

[26] Vgl. Dötsch/Pung, Organschaft, DB 2013, 307.

[27] Vgl. Dölker, in: Dölker/Ergenziger/Feldgen u.a., KStG, § 14 KStG, Rz. 19 (07.02.2020); Dötsch/Pung, in: D/P/M, KSt, § 14 KStG, Tz. 166 (August 2019).

[28] Vgl. Dölker, in: Dölker/Ergenziger/Feldgen u.a., KStG, § 14 KStG, Rz. 19 (07.02.2020).

[29] Vgl. § 14 Abs. 1 S. 1 Nr. 1 KStG.

[30] Vgl. § 14 Abs. 1 S. 1 Nr. 1 S. 1, 2 KStG.

[31] Vgl. R 14.2 KStH 2015, Bsp. 3.

[32] Vgl. Dölker, in: Dölker/Ergenziger/Feldgen u.a., KStG, § 14 KStG, Rz. 26 (02.02.2020); Heurung/Klübenspies, Organschaften, BB 2003, 2486; für die Additionsmethode: Dötsch/Pung, in: D/P/M, KSt, § 14 KStG, Tz. 267 (August 2016); Olbing, in: Streck, KStG, § 14 KStG, Rz. 51 (2018); Walter, in: Bott/Walter, KStG, § 14 KStG, Rz. 295.1 (August 2018); Prinz, Organschaftsbesteuerung, FR 2000, 1257; Orth, Zurechnungsfragen, Der Konzern 2005, 81 f.; für die Durchrechnungsmethode: Neumann, in: Gosch, KStG, § 14 KStG, Rz. 139 (2015).

[33] Vgl. § 14 Abs. 1 S. 1, S. 1 Nr. 3 KStG.

[34] Vgl. §§ 291 - 307 AktG.

[35] Vgl. § 17 Abs. 1 KStG.

[36] Vgl. BFH v. 30.07.1997, I R 7/97, BStBl. II 1998, 33, Leitsatz; Dötsch, in: D/P/M, KSt, § 17 KStG, Tz. 9 (Juni 2020); Boller/Hackemann, Prüfstand, IStR 2020, 44.

den.[37] Eine Nichtdurchführung des GAV innerhalb der ersten fünf Jahre ist schädlich mit Rückwirkung auf Beginn der Fünfjahresfrist.[38] Die Fünfjahresfrist bezieht sich allerdings nur auf den GAV und seine Durchführung, d.h. ein Fehlen einer sonstigen Tatbestandsvoraussetzung führt nur zu einer Versagung der Organschaft in dem betreffenden Jahr, nicht zu einer grundsätzlichen Versagung.[39]

Das Erfordernis des GAV anstelle einer steuerverfahrensrechtlichen Willenserklärung ist im Vergleich mit anderen Staaten eine Besonderheit des deutschen Steuerrechts.[40] Der GAV greift in die gesellschaftsrechtliche Organisation der beteiligten Gesellschaften ein und soll durch diesen Eingriff die steuerliche Verlustverrechnung der Gesellschaften untereinander rechtfertigen.[41]

2.2. Die steuerrechtlichen Folgen der Organschaft

2.2.1. Körperschaftsteuerliche Folgen

Wie oben beschrieben sind die Voraussetzungen für die Organschaft für Körperschaftsteuer und Gewerbesteuer identisch, die Rechtsfolgen unterscheiden sich jedoch. Bei der Körperschaftsteuer wird das Einkommen bei Organgesellschaft und Organträger zunächst unter Berücksichtigung der Sondervorschriften des § 15 KStG getrennt ermittelt und anschließend mit Ausnahme von 20/17 der ggf. erfolgten Ausgleichszahlungen an Minderheitsgesellschafter dem Organträger zugerechnet (sog. Zurechnungskonzept).[42] Die Organgesellschaft muss also 20/17 der Ausgleichszahlungen als eigenes Einkommen versteuern; dieser Betrag ergibt sich aus den Ausgleichszahlungen zzgl. dem Körperschaftsteuersatz von 15%, der auf die Ausgleichszahlungen aufgeschlagen wird.[43]

Grundsätzlich wird das Einkommen bei Organträger und -gesellschaft nach den üblichen Vorschriften des KStG ermittelt, abweichende Einkommensermittlungsvorschriften ergeben sich lediglich aus § 15 KStG.[44] Insbesondere werden Leistungsbeziehungen innerhalb des Organkreises aufgrund der separaten Einkommensermittlung nicht eliminiert.[45]

Die wichtigsten dieser Sondervorschriften sind, dass § 8b Abs. 1-6 KStG, also die Steuerbefreiung für verschiedene Einnahmen aus Kapitalbeteiligungen, nicht bei der Organgesellschaft anzuwenden sind, der Verlustabzug des § 10d EStG der Organgesellschaft versagt wird, was mangels der Möglichkeit für die Organgesellschaft, während der Organschaft eigene Verluste aufzubauen, faktisch zu einem Einfrieren der vororganschaftlichen Verluste für die Dauer der Organschaft führt, und die beteiligten Gesellschaften für Zwecke der Zinsschranke als ein Betrieb gelten und § 4h EStG bei der Organgesellschaft nicht anzuwenden ist.[46] Bei der Organgesellschaft ist die Gewinnabführung an den Organträger hinzuzurechnen bzw. der von dem Organträger geleistete Verlustausgleich abzuziehen.[47]

Bei dem Organträger ist die vereinnahmte Gewinnabführung vom Steuerbilanzgewinn bzw. -verlust abzuziehen.[48] Im Organkreis sind weiterhin verdeckte Gewinnausschüttungen möglich, etwa durch Unter- bzw. überpreislieferungen oder Zinsverlust, und stellen dann eine vorweggenommene Gewinnabführung dar.[49] Auch verdeckte Einlagen sind möglich und werden steuerlich nicht anders als außerhalb einer Organschaft behandelt.[50]

Das Organeinkommen ist gem. § 14 Abs. 1 S. 2 KStG erst für das Kalenderjahr zuzurechnen, in dem das Wirtschaftsjahr der Organgesellschaft endet, in dem der GAV wirksam wird, d.h. in dem alle Erfordernisse des GAV inkl. der Eintragung in das HR der Organgesellschaft erfüllt sind.[51]

Insgesamt ergeben sich durch die Organschaft mehrere Vorteile für die beteiligten Gesellschaften: Etwa können Gewinne und Verluste im Organkreis phasengleich ausgeglichen werden, die Zinsschranke gilt im Organkreis nicht und es muss im Organkreis keine Schachtelstrafe nach § 8b Abs. 5 KStG gezahlt werden.[52]

2.2.2. Gewerbesteuerliche Folgen

Bei einer gewerbesteuerlichen Organschaft gilt die Organgesellschaft als Betriebsstätte des Organträgers.[53] Dies nennt man die eingeschränkte Einheitstheorie, da die Unternehmen weiterhin selbstständig bleiben und ihre Gewerbeerträge einzeln ermitteln, die Gewerbesteuerpflicht allerdings

[37] Vgl. Dölker, in: Dölker/Ergenziger/Feldgen u.a., KStG, § 14 KStG, Rz. 29 (06.02.2020).

[38] Vgl. § 14 Abs. Nr. 3 KStG; R 14.5 (8) S. 1 Nr. 1 KStR 2015.

[39] Vgl. BFH v. 10.05.2017, I R 51/15, BStBl. II 2018, 30, Rn. 22-26; Heurung/Schmidt/Kraft, BB-Rechtsprechungsreport, BB 2018, 472.

[40] Vgl. Dölker, in: Dölker/Ergenziger/Feldgen u.a., KStG, § 14 KStG, Rz. 28 (06.02.2020); Boller/Hackemann, Prüfstand, IStR 2020, 43.

[41] Vgl. Müller-Gatermann, Organschaftsbesteuerung, in: Kley/Sünner/Willemsen, FS Ritter, 1997, 464; Scheunemann, Verlustberücksichtigung, IStR 2006, 146; van Lishaut, Reform, FR 2009, 1033; Cloer/van Lück, Verluste, DStRK 2019, 295.

[42] Vgl. R 7.1 KStR 2015; Krumm, in: Blümich, KStG, § 14 KStG, Rn. 201 (August 2018); Prinz/Ludwig, Gewinnabführungsvertrag, DB 2020, 1022.

[43] Vgl. § 16 S. 1 KStG; § 4 Abs. 5 Nr. 9 EStG i.V.m. § 8 Abs. 1 S. 1 KStG; Krumm, in: Blümich, KStG, § 16 KStG, Rn. 23 (August 2018).

[44] Vgl. Krumm, in: Blümich, KStG, § 14 KStG, Rn. 204 (August 2018); Dötsch/Pung, in: D/P/M, KSt, § 14 KStG, Tz. 692 (Juni 2019); Dölker, in: Dölker/Ergenziger/Feldgen u.a., KStG, § 14 KStG, Rz. 35 (06.02.2020).

[45] Vgl. van Lishaut, Reform, FR 2009, 1034; kritisch C. Witt, Konzernbesteuerung, FR 2009, 1046.

[46] Vgl. § 15 S. 1 Nr. 1 S. 1, Nr. 2 S. 1, Nr. 3 S. 1, 2 KStG; Krumm, in: Blümich, KStG, § 15 KStG, Rn. 10 (Februar 2019).

[47] Vgl. § 16 KStG; R 14.6 Abs. 1 S. 1 KStR 2015; Krumm, in: Blümich, KStG, § 14 KStG, Rn. 207 (August 2018); Dölker, in: Dölker/Ergenziger/Feldgen u.a., KStG, § 14 KStG, Rz. 36 (06.02.2020).

[48] Vgl. R 14.6 Abs. 1 S. 2 KStR 2015.

[49] Vgl. BFH v. 01.08.1984, I R 99/80, BStBl. II 1985, 18, Rz. 11; R 14.6 Abs. 4 KStR 2015.

[50] Vgl. BFH v. 15.03.2017, I R 67/15, BFH/NV 2017, 1276, Rn. 14; Neumann, in: Gosch, KStG, § 14 KStG, Rz. 401 (2015); Brink, in: Schnitger/Fehrenbacher, KStG, § 14 KStG, Rn. 738 (2018); Dölker, in: Dölker/Ergenziger/Feldgen u.a., KStG, § 14 KStG, Rz. 40 (06.02.2020).

[51] Vgl. § 14 Abs. 1 S. 2 KStG; Dölker, in: Dölker/Ergenziger/Feldgen u.a., KStG, § 14 KStG, Rz. 29 (06.02.2020).

[52] Vgl. § 15 S. 1 Nr. 2, 3 KStG; Kolbe, in: H/H/R, KStG, § 14 KStG, Anm. 10; Prinz, Organschaftsrecht, Beihefter zu DStR 2010, 70; Ismer, Gruppenbesteuerung, DStR 2012, 822; Müller/Detmering/Lieber, Organschaft, 11. Aufl. 2020, Rz. 9; Prinz/Ludwig, Gewinnabführungsvertrag, DB 2020, 1022.; Prinz, in: Prinz/Witt, Organschaft, 2. Aufl., Rz. 1.21.

[53] Vgl. § 2 Abs. 2 S. 2 GewStG.

nur dem Organträger zugerechnet wird.[54] Hierdurch wird die Gemeinde, in deren Bezirk das Unternehmen der Organgesellschaft liegt, nicht in ihrem Anspruch auf Gewerbesteuer benachteiligt, auch wenn sich das Unternehmen des Organträgers im Bezirk einer anderen Gemeinde befindet.[55] Hat die Organgesellschaft mehrere Betriebsstätten, gelten diese jeweils als Betriebsstätten des Organträgers.[56]

Auch in der Gewerbesteuer darf die Organgesellschaft ihren vor der Organschaft entstandenen gewerbesteuerlichen Verlustvortrag nicht während der Organschaft nutzen, um ihren maßgebenden Gewerbeertrag zu kürzen.[57] Für Dividendeneinnahmen gilt die Sonderregelung des § 7a GewStG, wonach eine von der Organgesellschaft bezogene Dividende nicht nach § 9 Nr. 2a, 7, 8 GewStG gekürzt werden soll und für damit unmittelbar zusammenhängende Aufwendungen § 8 Nr. 1 GewStG nicht anzuwenden ist.[58] Durch die Betriebsstättenfiktion der Gewerbesteuer wird der Gewerbesteuermessbetrag auf die einzelnen Gemeinden aufgeteilt.[59] Dabei kann es durch den Zerlegungsmaßstab lt. § 29 Abs. 1 Nr. 1 GewStG und die unterschiedlichen Gewerbesteuerhebesätze in den verschiedenen Gemeinden zu Mehr- oder Minderbelastungen durch die Organschaft kommen.[60]

2.3. Verlustabzugsverbot des § 14 Abs. 1 S. 1 Nr. 5 KStG

Gem. § 14 Abs. 1 S. 1 Nr. 5 KStG bleiben negative Einkünfte von Organträger oder Organgesellschaft von der Steuer unberücksichtigt, soweit sie in einem ausländischen Staat im Rahmen der Besteuerung von Organträger, Organgesellschaft oder einer anderen Person berücksichtigt werden.[61] Das Ziel dieser Vorschrift ist die Verhinderung einer doppelten Verlustnutzung und des Imports von Auslandsverlusten.[62] Diese Vorschrift ist also grds. bei Auslandssachverhalten zu berücksichtigen, da eine ausländische Verlustnutzung von den ausländischen Steuervorschriften und somit von dem Einzelfall abhängt, sie birgt allerdings große Unsicherheiten. Es wird jedoch vermutet, dass die Vorschrift bisher seitens der Finanzverwaltung nur selten oder sogar gar nicht angewandt wurde.[63]

Der persönliche Anwendungsbereich erstreckt sich auf alle Organgesellschaften und Organträger, eine doppelte Ansässigkeit ist nicht notwendig.[64] Unklar ist, ob die Vorschrift auch auf Organträger in der Rechtsform einer PersG anzuwenden ist, da eine PersG kein eigenes Steuersubjekt ist und somit auch keine negativen Einkünfte haben kann.[65] Möglich ist jedoch, dass die Formulierung „negative Einkünfte“ in dieser Hinsicht lediglich eine sprachliche Ungenauigkeit darstellt, trotz derer die Vorschrift auf Personengesellschaften als Organträger angewandt werden kann.[66] Dieser Ansicht ist sich m.E. anzuschließen, da nicht ersichtlich ist, weshalb der Gesetzesgeber Personengesellschaften von dieser Regelung ausnehmen wollte und insb. das Abstellen auf die Besteuerung einer anderen Person auf den Mitunternehmer der PersG abzielt.

Laut Entscheidung des BFH liegen beim Organträger nur dann negative Einkünfte vor, wenn nach Zurechnung des Einkommens der Organgesellschaft Verluste verbleiben.[67] Diese Entscheidung hat Spekulationen aus der Literatur, dass Organträger und Organgesellschaft separat betrachtet werden müssten (sog. „Stand-Alone-Betrachtung“), ein Ende gesetzt.[68] Der Begriff „negative Einkünfte“ bezieht sich aber wohl nur auf die Einkünfte aus Gewerbebetrieb, sodass etwa bei einer natürlichen Person als Organträger die Einkünfte aus anderen Einkunftsarten irrelevant sind.[69]

Die ausländischen Einkünfte richten sich nach den Vorschriften im Ausland, d.h. es muss sowohl im Inland als auch im Ausland ein negativer Einkommenssaldo abgezogen werden und dieser muss sich aus denselben Posten zusammensetzen.[70] Einzelne im Ausland abgezogene Betriebsausgaben oder Posten sind also auch im Inland abziehbar, wenn die Gesellschaft insgesamt positive Einkünfte erzielt hat.[71] Anderenfalls müsste im Normtext die Rede von Einkünfteteilen anstelle von Einkünften sein.[72] Der Begriff der Verlustberücksichtigung ist nicht eindeutig, es muss aber lt. Gesetzesbe-

[54] Vgl. BFH v. 18.09.1996, I R 44/95, BStBl. II 1997, 181, Rz. 9; BFH v. 22.04.1998 - I R 109/97, BStBl. II 1998, 748, Rz. 9; BFH v. 18.05.2011, X R 4/10, BStBl. II 2011, 887, Rn. 38; BFH v. 17.12.2014, I R 39/14, BStBl. II 2015, 1052, Rn. 8; BFH v. 07.09.2016, I R 9/15, BFH/NV 2017, 485, Rn. 16; Prinz/Ludwig, Gewinnabführungsvertrag, DB 2020, 1022.

[55] Vgl. Leicht, Organschaft, in: Verlag C.H.BECK, Bilanzrechtslexikon, Rn. 39 (01.01.2020); Keß, in: L/S, GewStG, § 2 GewStG, Rz. 3510 f. (Februar 2020);.

[56] Vgl. Keß, in: L/S, GewStG, § 2 GewStG, Rz. 3523 (Februar 2020).

[57] Vgl. § 10a S. 3 GewStG.

[58] Vgl. § 7a Abs. 1 S. 1 GewStG.

[59] Vgl. § 28 Abs. 1 S. 1 GewStG.

[60] Vgl. Grotherr, Bedeutung, BB 1993, 1999.

[61] Vgl. § 14 Abs. 1 S. 1 Nr. 5 KStG.

[62] Vgl. BT-Drucks. 17/10774, 20; Dölker, in: Dölker/Ergenziger/Feldgen u.a., KStG, § 14 KStG, Rz. 41 (06.02.2020); Prinz/Ludwig, Gewinnabführungsvertrag, DB 2020, 1023.

[63] Vgl. Walter, in: Bott/Walter, KStG, § 14 KStG, Rz. 958 (Februar 2020).

[64] Vgl. Kraft/Hohage, Abzugsverbot, Ubg 2019, 306.

[65] Vgl. Walter, in: Bott/Walter, KStG, § 14 KStG, Rz. 958 (Februar 2020); ablehnend: Benecke/Schnitger, änderungen, IStR 2013, 147; Kraft/Hohage, Abzugsverbot, Ubg 2019, 306 f.; befürwortend: Krumm, in: Blümich, KStG, § 14 KStG, Rn. 220 (August 2018); Dötsch/Pung, in: D/P/M, KSt, § 14 KStG, Tz. 654 (Oktober 2018); Schaden/Polatzky, Verlustausgleichsbeschränkung, IStR 2013, 133; von Freeden/Liekenbrock, Zinsabzugsbeschränkung, DB 2013, 1691; Ritzer/Aichberger, Anti-Double-Dip-Regelung, Der Konzern 2013, 603 f.; Wacker, I R 92/12 (Anm.), IStR 2017, 288; Prinz, Sonderbetriebsausgaben, GmbHR 2017, 555 f.

[66] Vgl. Neumann, in: Gosch, KStG, § 14 KStG, Rn. 480 (2015); Kolbe, in: H/H/R, KStG, § 14 KStG, Anm. 271 (Januar 2015); Brink, in: Schnitger/Fehrenbach, KStG, § 14 KStG, Rn. 969e (2018); Benecke/Schnitger, änderungen, IStR 2013, 147; Jesse, Neuregelungen, FR 2013, 638.

[67] Vgl. BFH v. 12.10.2016, I R 92/12, BFH/NV 2017, 685, Rn. 49; gl. A. Meilicke, Neuregelung, DB 2002, 914.

[68] Vgl. Benecke/Schnitger, änderungen, IStR 2013, 145.

[69] Vgl. Benecke/Schnitger, änderungen, IStR 2013, 146.

[70] Vgl. Meilicke, Neuregelung, DB 2002, 916.

[71] Vgl. BFH v. 12.10.2016, I R 92/12, BFH/NV 2017, 685, Rn. 46; Neumann, in: Gosch KStG, § 14 KStG, Rn. 480 (2015); Rödder/Liekenbrock, in: Rödder/Herlinghaus/Neumann, KStG, § 14 KStG, Rz 449 (2015); Schneider/Schmitz, Verlustberücksichtigung, GmbHR 2013, 282 f.; Schaden/Polatzky, Verlustausgleichsbeschränkung, IStR 2013, 136; Prinz, Abwehrgesetzgebung, DB 2018, 1619.

[72] Vgl. Kolbe, in: H/H/R, KStG, § 14 KStG, Anm. 276 (Januar 2015); Kraft/Hohage, Abzugsverbot, Ubg 2019, 308; zweifelnd Benecke/Schnitger, änderungen, IStR 2013, 151.

gründung ein Abzug vom ausländischen Einkommen vorliegen, d.h. eine Berücksichtigung etwa im Rahmen eines Progressionsvorbehalts soll nicht ausreichen.[73]

Die Vorschrift führt zu kritischen Einzelfallkonstellationen, so kann eine inländische Kapitalgesellschaft im Normalfall die Verluste ihrer ausländischen Betriebsstätte (unter Berücksichtigung von § 2a EStG) bei der inländischen Besteuerung abziehen, dies wird jedoch nicht mehr möglich sein, wenn die Kapitalgesellschaft Organträger oder Organgesellschaft ist, selbst wenn die ausländische Betriebsstätte außerhalb des Organkreises gehalten wird.[74] Insgesamt ist auch ein Verstoß gegen das im GG kodifizierte Leistungsfähigkeitsprinzip denkbar, da die Verluste endgültig nicht abziehbar werden und nicht etwa nur zeitlich vorzutragen sind, während die spiegelbildlichen Gewinne der inländischen Besteuerung unterliegen.[75]

Es steht weiterhin zur Diskussion, ob die Vorschrift auf die Gewerbesteuer anzuwenden ist. § 2 Abs. 2 S. 2 GewStG bezieht sich nur auf die Voraussetzungen für die körperschaftsteuerliche Organschaft, nicht auf die Rechtsfolgen.[76] Das Verlustabzugsverbot ist jedoch nach h.M. als Gewinn- bzw. Einkünfteermittlungsvorschrift und nicht als Einkommensermittlungsvorschrift anzusehen, was dazu führt, dass der Verweis in § 7 GewStG auch für § 14 Abs. 1 S. 1 Nr. 5 KStG gilt.[77] Diese Sicht wird auch vom BFH unterstützt, da die Vorschrift ansonsten aufgrund der Unmöglichkeit des negativen Einkommens bei der Organgesellschaft für Organgesellschaften ins Leere liefe, zudem entspricht diese Sichtweise dem Gesetzeswortlaut („negative Einkünfte“).[78] Da es in der Anlage zur Ermittlung der Einkünfte aus Gewerbebetrieb der amtlichen Körperschaftsteuerformulare ein Feld für die Hinzurechnung von Verlusten nach § 14 Abs. 1 S. 1 Nr. 5 KStG gibt, ist die Vorschrift nach Meinung der Finanzverwaltung offensichtlich eine Einkünfteermittlungsvorschrift und mangels Korrekturvorschriften im Gewerbesteuerrecht auch auf die Gewerbesteuer anzuwenden.[79] Gegen die tatsächliche Anwendung gibt es allerdings noch weitere Argumente, etwa, dass in der Gewerbesteuer das Territorialitätsprinzip gilt, wonach eine inländische Betriebsstätte für die Besteuerung entscheidend ist und somit ausländische Merkmale ggf. außen vor zu lassen seien.[80] Zudem gibt es im Regelfall keine der Gewerbesteuer entsprechende ausländische Steuer, wodurch eine doppelte Verlustberücksichtigung bei der Gewerbesteuer grundsätzlich nicht möglich ist und somit auch keine Hinzurechnung der ausländischen Verluste nötig ist.[81]

Wird die Vorschrift aufgrund einer in einem EU- oder EWR-Staat abgezogenen Steuer angewandt, liegt mit hoher Wahrscheinlichkeit auch eine Unionsrechtswidrigkeit vor, da der EuGH eine nahezu identisch formulierte Vorschrift des britischen Steuerrechts als eine nicht gerechtfertigte Einschränkung der Niederlassungsfreiheit beurteilt hat.[82] Insgesamt wird die Vorschrift aufgrund ihres weitreichenden Anwendungsbereichs und der unklaren Formulierungen scharf kritisiert und eine änderung oder Abschaffung gefordert.[83]

3. Inbound-Sachverhalt

3.1. EU-/EWR-Staat

3.1.1. Ausländisches Mutterunternehmen mit inländischer Betriebsstätte

Rechtsstand

Der aus organschaftlicher Sicht womöglich unproblematischste Inbound-Fall ist der eines ausländischen Mutterunternehmens mit einer inländischen Betriebsstätte, welcher die Beteiligung an der inländischen Tochtergesellschaft zugeordnet ist.

In diesem Fall wird für die Zwecke dieser Arbeit unterstellt, dass die persönlichen Voraussetzungen der inländischen Organgesellschaft erfüllt sind, ebenso wie das Erfordernis der finanziellen Eingliederung. Bei diesen beiden Voraussetzungen kommt es zu keinen zusätzlichen Schwierigkeiten aufgrund der Fallkonstellation. Zu prüfen sind also insb. das GAV-Erfordernis sowie die persönlichen Voraussetzungen des Organträgers. Sind diese Voraussetzungen erfüllt, liegt sowohl eine gewerbe- als auch eine körperschaftsteuerliche Organschaft vor.

Für den Organträger gibt es seit der sog. kleinen Organschaftsreform nicht mehr das Erfordernis der unbeschränkten Steuerpflicht, wie es vor der Reform der Fall war, d.h. es sind nicht mehr Sitz oder Ort der GL im Inland notwendig.[84] Stattdessen muss die Beteiligung während der

[73] Vgl. BT-Drucks. 17/10774, 20; gl. A. Meilicke, Neuregelung, DB 2002, 914; Benecke/Schnitger, änderungen, IStR 2013, 150.

[74] Vgl. Dötsch/Pung, Organschaft, DB 2013, 312 f.

[75] Vgl. Art. 3 GG; Meilicke, Neuregelung, DB 2002, 916; Benecke/Schnitger, änderungen, IStR 2013, 151.

[76] Vgl. § 2 Abs. 2 S. 2 GewStG; Schneider/Schmitz, Verlustberücksichtigung, GmbHR 2013, 284; Kempf/Loose/Oskamp, Verlustbeschränkung, DStR 2018, 2512.

[77] Vgl. Frotscher, in: F/D, KStG, § 2 GewStG, Rz. 75c (Januar 2018) ; Benecke/Schnitger, änderungen, IStR 2013, 147 f.; Jesse, Neuregelungen, FR 2013, 638; Kempf/Loose/Oskamp, Verlustbeschränkung, DStR 2018, 2513.

[78] Vgl. BFH v. 12.10.2016, I R 93/12, BFH/NV 2017, 685, Rn. 45.

[79] Vgl. Anlage GK der Formulare zur Körperschaftsteuererklärung 2019, Zeile 163; Frotscher, in: F/D, KStG, § 2 GewStG, Rz. 75c (Januar 2018); Benecke/Schnitger, änderungen, IStR 2013, 148 (Anm. Benecke); Kempf/Loose/Oskamp, Verlustbeschränkung, DStR 2018, 2513.

[80] Vgl. Wagner/Liekenbrock, Verlustberücksichtigung, Ubg 2013, 141; Benecke/Schnitger, änderungen, IStR 2013, 148 (Anm. Schnitger); Ritzer/Aichberger, Anti-Double-Dip-Regelung, Der Konzern 2013, 604; Kempf/Loose/Oskamp, Verlustbeschränkung, DStR 2018, 2512.

[81] Vgl. Frotscher, in: F/D, KStG, § 2 GewStG, Rz. 75c (Januar 2018); Löwenstein/Maier, Verlustnutzung, IStR 2002, 192 f.; Endres/Thies, Verlustnutzung, RIW 2002, 278 f.; Jesse, Neuregelungen, FR 2013, 638; Ritzer/Aichberger, Anti-Double-Dip-Regelung, Der Konzern 2013, 604; IDW, Zweifelsfragen, Ubg 2014, 216; Kempf/Loose/Oskamp, Verlustbeschränkung, DStR 2018, 2512.

[82] Vgl. EuGH v. 06.09.2012, C-18/11, Rs. „Philips Electronics UK“, EU:C:2012:532, Rn. 20 u. 35; Krumm, in: Blümich, KStG, § 14 KStG, Rn. 42 (August 2018); Schober, in: Musil/Weber-Grellet, Steuerrecht, § 14 KStG, Rn. 302 (2019); Meilicke, Neuregelung, DB 2002, 916.

[83] Vgl. Meilicke, Neuregelung, DB 2002, 917; Schwenke, Organschaft, ISR 2013, 46; Kessler/Arnold, Organschaft, IStR 2016, 230.

[84] Vgl. Gesetz zur änderung und Vereinfachung der Unternehmensbesteuerung und des steuerlichen Reisekostenrechts v. 20.02.2013, BGBl. I 2013, 285; BT-Drucks. 17/10774, 18.

gesamten Dauer der Organschaft einer inländischen Betriebsstätte zuzuordnen sein, wodurch sichergestellt wird, dass Deutschland mindestens im Rahmen einer beschränkten Steuerpflicht Zugriff auf das Steuersubstrat hat.[85] Dementsprechend kann der Organträger selbst Sitz und Ort der GL in einem EU-/EWR-Staat haben, ohne dass dies problematisch ist; auch ein doppelt ansässiger Organträger oder eine nach deutschem Recht gegründete Gesellschaft mit Ort der GL im Ausland sind unproblematisch.[86] Die Einkünfte der Betriebsstätte müssen allerdings sowohl nach deutschem als auch nach DBA-Recht der inländischen Besteuerung unterliegen, wodurch bei etwaigen Definitionsdifferenzen die Entstehung von sog. weißen Einkünften, d.h. Einkünften, die gar nicht besteuert werden, vermieden werden soll.[87]

Zuordnung zu einer inländischen Betriebsstätte

Die Kriterien, nach denen die Beteiligung einer Betriebsstätte zugeordnet werden soll, sind nicht im Gesetz kodifiziert.[88] Es herrscht allerdings die Meinung vor, dass eine Zuordnung aufgrund eines funktionalen Zusammenhangs zur Geschäftstätigkeit der Betriebsstätte zu erfolgen hat, etwa auf Basis der Betriebsstättengewinnabgrenzungsverordnung.[89] Liegt kein funktionaler Zusammenhang zu einer einzelnen Betriebsstätte vor, führt nach Meinung der Finanzverwaltung die Zentralfunktion des Stammhauses zu einer Zuordnung.[90] Diese grundsätzliche Zuordnung zum Stammhaus kann bei einem ausländischen beherrschenden Unternehmen mit inländischer Betriebsstätte zu einem Problem werden, da bei ihm ein Fehlen der funktionalen Zuordnung zu einer Versagung der Organschaft führt.[91] Nach anderer Meinung ist die Zuordnung zur Betriebsstätte für die inländische Besteuerung lediglich nach der Zuordnung zum gewillkürten Betriebsvermögen vorzunehmen,[92] was für die Zwecke des § 14 Abs. 1 KStG jedoch nicht zielführend wäre, da das Erfordernis der Zuordnung dann identisch mit dem der finanziellen Eingliederung wäre.[93] Voraussetzung für die Zuordnung zu einer Betriebsstätte ist allerdings, dass die Beteiligung grundsätzlich dem Betriebsvermögen des Organträgers zuzuordnen ist.[94]

Ist das Mutterunternehmen eine ausländische vermögensverwaltende Holding, kann die funktionale Zuordnung der Beteiligung zur Betriebsstätte nicht erfolgen, da die Voraussetzungen einer Betriebsstätte i.S.d. § 12 AO nicht erfüllt sind.[95] Bei einer geschäftsleitenden Holding muss für die Zuordnung die gewerbliche Tätigkeit von der inländischen Betriebsstätte ausgeübt werden, sodass in diesem Fall auch ein inländischer Ort der GL vorläge.[96] Zu beachten ist weiterhin, dass die Zuordnung bei einem ausländischen Mutterunternehmen nach h.M. zu einer einzigen Betriebsstätte erfolgen muss, d.h. die Beteiligung, die die finanzielle Eingliederung vermittelt, muss einer inländischen Betriebsstätte zuzuordnen sein, sodass eine Zusammenrechnung der Beteiligungen mehrerer Betriebsstätten nicht möglich ist.[97]

Aufgrund des Erfordernisses, dass die der Betriebsstätte zuzurechnenden Einkünfte auch nach dem anzuwendenden DBA der inländischen Besteuerung unterliegen müssen, muss die Beteiligung auch nach DBA-Recht der Betriebsstätte zuzuordnen sein, was i.d.R. ebenfalls nach dem tatsächlichen funktionalen Zusammenhang erfolgt.[98] Das Erfordernis des funktionalen Zusammenhangs ist nach h.M. erfüllt, wenn Nebenerträge der Tätigkeit der Betriebsstätte vorliegen[99] und die Beteiligung tatsächlich eine Funktion für die Betriebsstätte ausübt,[100] es müssen also zwischen Organgesellschaft und Betriebsstätte wesentliche Liefer- bzw. Leistungsbeziehungen bestehen.[101] Irrelevant ist, ob die Stimmrechte von der Betriebsstätte oder dem Stammhaus aus ausgeübt werden.[102]

[85] Vgl. § 14 Abs. 1 S. 1 Nr. 2 S. 1, 4 KStG; OFD Frankfurt/M. v. 12.11.2019, S 2770 A - 55- St 55, DB 2020, 257.

[86] Vgl. Gesetz zur änderung und Vereinfachung der Unternehmensbesteuerung und des steuerlichen Reisekostenrechts v. 20.02.2013, BGBl. I 2013, 285.; OFD Frankfurt/M. v. 12.11.2019, S 2770 A - 55- St 55, DB 2020, 257; Schnitger, Fragestellungen, IStR 2013, 84; Blumenberg/Hundeshagen, Gestaltungsfragen, in: Prinz/Witt, Organschaft, 2. Aufl. 2019, 373, Rz. 8.81.

[87] Vgl. § 14 Abs. 1 S. 1 Nr. 2 S. 7 KStG; Keß, in: L/S, GewStG, § 2 GewStG, Rz. 3614 (Oktober 2017); Dötsch/Pung, Organschaft, DB 2013, 308.

[88] Vgl. IDW, Stellungnahme, https://www.idw.de/blob/83816/49b57b68f1bd41c9a34dc510db3f963c/down-bmf-kstr-e2015-anl2-data.pdf, abgerufen am 30.05.2020, zu R 14.2 und R 14.3 KStR-E 2015, S. 7.

[89] Vgl. § 7 Abs. 1 S. 1 f. BsGaV; BFH v. 26.02.1992, I R 85/91, BStBl. II 1992, 937, Rz. 21; BMF v. 24.12.1999, IV B 4 - S 1300 - 111/99, BStBl. I 1999, 1076, Tz. 2.4; Dölker, in: Dölker/Ergenziger/Feldgen u.a., KStG, § 14 KStG, Rz. 22 (07.02.2020); Prinz/Ludwig, Gewinnabführungsvertrag, DB 2020, 1023.

[90] Vgl. BMF v. 24.12.1999, IV B 4 - S 1300 - 111/99, BStBl. I 1999, 1076, Tz. 2.4.

[91] Vgl. Kolbe, in: H/H/R, KStG, § 14 KStG, Anm. 189 (September 2016).

[92] Vgl. Keß, in: L/S, GewStG, § 2 GewStG, Rz. 3616 (Oktober 2017); a.A. Schirmer, Holding, GmbHR 2013, 797.

[93] Vgl. Schirmer, Holding, GmbHR 2013, 797.

[94] Vgl. BMF v. 24.12.1999, IV B 4 - S 1300-111/99, BStBl. I 1999, 1076, Tz. 2.4; Kolbe, in: H/H/R, KStG, § 14 KStG, Anm. 189 (September 2016); Dötsch/Pung, in: D/P/M, KSt, § 14 KStG, Tz. 182 (Juni 2020); Schirmer, Zuordnung, FR 2013, 607 f.; Goebel/Ungemach, Neuregelungen, NWB 2013, 596 f.; Weigert/Strohm, Organschaft, Der Konzern 2013, 267; Jesse, Neuregelungen, FR 2013, 634 f.

[95] Vgl. Kolbe, in: H/H/R, KStG, § 14 KStG, Anm. 189 (September 2016); Schulze zur Wiesche, Organschaft, DStZ 2013, 627; Goebel/Ungemach, Neuregelungen, NWB 2013, 599; Weigert/Strohm, Organschaft, Der Konzern 2013, 267; Schirmer, Holding, GmbHR 2013, 798 f.

[96] Vgl. Kolbe, in: H/H/R, KStG, § 14 KStG, Anm. 189 (September 2016); Schulze zur Wiesche, Organschaft, DStZ 2013, 627; Schirmer, Holding, GmbHR 2013, 799.

[97] Vgl. Kolbe, in: H/H/R, KStG, § 14 KStG, Anm. 189 (September 2016); Keß, in: L/S, GewStG, § 2 GewStG, Rz. 3618 (Oktober 2017); Müller, in: M/S/O, KStG, § 14 KStG, Rz. 84 (2019).

[98] Vgl. BFH v. 19.12.2007, I R 66/06, BStBl. II 2008, 510, Rn. 17; BMF v. 16.4.2010, IV B 2 - S 1300/09/10003, BStBl. I 2010, 354, Tz. 2.2.4.1; Keß, in: L/S, GewStG, § 2 GewStG, Rz. 3617 (Oktober 2017); Güroff, in: Glanegger/Güroff, GewStG, § 2 GewStG, Rz. 495 (2017); Dötsch/Pung, Organschaft, DB 2013, 307 f; Schirmer, Holding, GmbHR 2013, 797; Herbener, Holding, in: Prinz/Witt, Organschaft, 2. Aufl. 2019, 676, Rz. 17.11; Dötsch/Pung, Zuordnung, DB 2019, 147.

[99] Vgl. BFH v. 19.12.2007, I R 66/06, BStBl. II 2008, 510, Rn. 17; Schirmer, Holding, GmbHR 2013, 799; Dötsch, Organträger-Personengesellschaft, DB 2014, 1216.

[100] Vgl. Schirmer, Holding, GmbHR 2013, 797.

[101] Vgl. Keß, in: L/S, GewStG, § 2 GewStG, Rz. 3617 (Oktober 2017).

[102] Vgl. Kolbe, in: H/H/R, KStG, § 14 KStG, Anm. 189 (Januar 2015); Keß, in: L/S, GewStG, § 2 GewStG, Rz. 3618 (Oktober 2017); Stangl/Brühl, Organschaftsreform, Der Konzern 2013, 80.

Besonders zu beachten ist der Fall einer PersG mit ausländischen Mitunternehmern als beherrschende Gesellschaft. In diesem Fall führt das zivilrechtliche Gesamthandseigentum nach Meinung der Finanzverwaltung zu einer inländischen Betriebsstätte.[103] Die Zuordnung der Beteiligung an der Tochtergesellschaft zur inländischen Betriebsstätte hängt dabei neben der der Ansässigkeit der Tochtergesellschaft auch von der der Mitunternehmer ab.[104]

Bei einer mittelbaren Beteiligung an der Organgesellschaft muss nach h.M. lediglich die Beteiligung an der vermittelnden Gesellschaft während der Dauer der Organschaft einer inländischen Betriebsstätte zuzuordnen sein,[105] die Zuordnung bei der vermittelnden Gesellschaft ist damit nicht relevant.[106] Die vermittelnde Gesellschaft selbst muss nicht die Voraussetzungen einer Organgesellschaft erfüllen, kann also auch im Ausland ansässig sein, was einem unionsrechtlichen Erfordernis entspricht, dass auch eine im Gemeinschaftsgebiet ansässige Tochtergesellschaft eine Konzernbesteuerung mit einer im Inland ansässigen Enkelgesellschaft vermitteln können muss.[107] Insgesamt ist die Regelung insofern ungenau formuliert, als dass lediglich das Organeinkommen, nicht aber alle der Betriebsstätte zuzurechnenden Einkünfte nach inländischen Vorschriften ebenso wie nach DBA der deutschen Besteuerung unterliegen müssen, da eine andere Auslegung zu nicht stimmigen Ergebnissen führen würde.[108]

Die Rechtsfolge der inländischen Zuordnung ist gem. § 14 Abs. 1 S. 1 Nr. 2 S. 6 KStG die Zurechnung des Einkommens der Organgesellschaft zu der jeweiligen Betriebsstätte, sodass es gem. § 49 Abs. 1 Nr. 2 a) EStG der beschränkten Steuerpflicht unterliegt und beim Organträger zu versteuern ist.[109]

Gewinnabführungsvertrag

Das zweite potenzielle Problemfeld ist der GAV. Da dieser ein Institut des deutschen Gesellschaftsrechts ist, sind Probleme beim Abschluss mit einem ausländischen Mutterunternehmen theoretisch denkbar.[110] Tatsächlich bestimmt sich der GAV aber nach h.M. nach dem Gesellschaftsstatut der beherrschten, also der abführenden Gesellschaft, da tiefer in ihre Autonomie eingegriffen wird.[111] Dementsprechend gilt hier das Recht der Organgesellschaft, also das deutsche Gesellschaftsrecht, der Abschluss mit einem ausländischen Mutterunternehmen wird deshalb für zulässig gehalten.[112] Hiervon abweichend wurde in einer Einzelmeinung zusätzlich die Voraussetzung formuliert, dass das ausländische Unternehmen im deutschen Konzernrecht als Unternehmen gilt.[113] Das ausländische Recht muss allerdings auch die Pflichten des herrschenden Unternehmens zulassen, was in der Regel der Fall ist.[114] Der BFH hat diese Ansicht bestätigt, für die steuerliche Anerkennung jedoch - ebenso wie bei einer inländischer Organschaft - die exakte Befolgung der formellen und materiellen Voraussetzungen während der gesamten Laufzeit des GAV gefordert, um eine Vergleichbarkeit herzustellen.[115] Nach anderer Meinung sind nur die Vorschriften anzuwenden, die dem Schutz der Tochtergesellschaft, ihrer Gläubiger und ihrer Gesellschafter dienen.[116] Teilweise ist man auch der Meinung, ein GAV mit einem ausländischen beherrschenden Unternehmen sei komplett auszuschließen.[117] Da diese Meinungen in der Minderheit sind, ist m.E. der Ansicht des BFH zu folgen.

3.1.2. Ausländisches Mutterunternehmen ohne inländische Betriebsstätte

Einbeziehung des Mutterunternehmens in den Organkreis

Nach aktuellem Rechtsstand muss, wie oben ausgeführt, das Mutterunternehmen über eine inländische Betriebsstätte verfügen, welcher für die Dauer der Organschaft die Beteiligung an der Organgesellschaft zugerechnet wird.[118] Eine

[103] Vgl. OFD Karlsruhe v. 16.01.2014, S 2770/52/2 - St 221, FR 2014, 434; Benecke/Schnitger, änderungen, IStR 2013, 153.

[104] Genauer hierzu vgl. Wacker, I R 58/15 (Anm.), FR 2018, 562; Dötsch/Pung, Zuordnung, DB 2019, 146 f.

[105] Vgl. § 14 Abs. 1 S. 1 Nr. 2 S. 5 KStG; Kolbe, in: H/H/R, KStG, § 14 KStG, Anm. 191 (September 2016); Dötsch/Pung, in: D/P/M, KSt, § 14 KStG, Tz. 182 (Juni 2019); Dölker, in: Dölker/Ergenziger/Feldgen u.a., KStG, § 14 KStG, Rz. 26 (02.02.2020); Dötsch/Pung, Organschaft, DB 2013, 307; Jesse, Neuregelungen, FR 2013, 635; a.A. Schirmer, Zuordnung, FR 2013, 608.

[106] Vgl. Kolbe, in: H/H/R, KStG, § 14 KStG, Anm. 191 (September 2016); Benecke/Schnitger, änderungen, IStR 2013, 154.

[107] Vgl. BFH v. 02.11.1977, I R 143/75, BStBl. II 1978, 74, Rn. 16-18; EuGH v. 27.11.2008, C-418/07, Rs. „Société Papillon/Ministère du Budget, des Comptes publics et de la Fonction publique“, EU:C:2008:659, Rn. 63; EuGH v. 12.06.2014, C-39/13, C-40/13, C-41/13, Rs. „SCA Group Holding u.a.“, EU:C:2014:1758, Rn. 43; H 14.2 „Finanzielle Eingliederung“ unter „Mittelbare Beteiligung“ KStH 2015; Neumann, in: Gosch, KStG, § 14 KStG, Rz. 150 (2015); Dötsch/Pung, in: D/P/M, KSt, § 14 KStG, Tz. 263 (Juni 2018); Ismer, Gruppenbesteuerung, DStR 2012, 823; Dötsch/Pung, Organschaft, DB 2013, 308.

[108] Vgl. Rödder, Organschaftsreform, Ubg 2012, 722.

[109] Vgl. § 14 Abs. 1 S. 1, S. 1 Nr. 2 S. 6 KStG; Kolbe, in: H/H/R, KStG, § 14 KStG, Anm. 194 (September 2016).

[110] Vgl. Koch, in: Hüffer/Koch, AktG, § 291 AktG, Rn. 4a (2020).

[111] Vgl. BFH v. 07.12.2011, I R 30/08, BStBl. II 2012, 507, Rn. 21; Bärwaldt/Schabacker, Unternehmensverträge, AG 1998, 186; Maul, Vollstreckungsfragen, AG 1998, 407 f.; Picot/Land, Unternehmenskauf, DB 1998, 1606; Scheunemann, Verlustberücksichtigung, IStR 2006, 146; van Lishaut, Reform, FR 2009, 1032; Frotscher, Organschaft, IStR 2011, 702; Glahe, Organschaft, IStR 2012, 131; Hoene, Gewinnabführungsvertrag, IStR 2012, 463; Kessler/Arnold, Organschaft, IStR 2016, 230; Koehler, Gewinnabführungsvertrag, Der Konzern 2018, 326; Krüger/Epe, Fragen, in: Prinz/Witt, Organschaft, 2. Aufl. 2019, 928 f., Rz. 26.4; a.A., wonach Gesellschaftsstatute beider Gesellschaften gelten sollen: Einsele, Unternehmen, ZGR 1996, 50.

[112] Vgl. Hommelhoff, in: Lutter/Hommelhoff, GmbHG, Anhang zu § 13, Rz. 96 (2020); Koehler, Gewinnabführungsvertrag, Der Konzern 2018, 326; Prinz/Ludwig, Gewinnabführungsvertrag, DB 2020, 1023; Weinberger, Gewinnabführungsvertrag, Der Konzern 2020, 63.

[113] Vgl. Frotscher, Organschaft, IStR 2011, 701.

[114] Vgl. Hommelhoff, in: Lutter/Hommelhoff, GmbHG, Anhang zu § 13, Rz. 96 (2020); Glahe, Organschaft, IStR 2012, 131; Prinz/Ludwig, Gewinnabführungsvertrag, DB 2020, 1023.

[115] Vgl. BFH v. 07.12.2011, I R 30/08, BStBl. II 2012, 507, Rn. 21 f. u. 24.

[116] Vgl. OLG Stuttgart v. 05.06.2013, 20 W 6/10, NZG 2013, 897, Rn. 131; Liebscher, in: Fleischer/Goette, GmbHG, Anhang zu § 13, Rn 1236 (2018); Krüger/Epe, Fragen, in: Prinz/Witt, Organschaft, 2. Aufl. 2019, 929 f., Rz. 26.7.

[117] Vgl. Balmes/Brück/Ribbrock, Organschaftsbesteuerung, BB 2005, 969.

[118] Vgl. § 14 Abs. 1 Nr. 2 S. 4 KStG.

Organschaft zwischen beherrschendem und beherrschtem Unternehmen kann also nicht zustande kommen, wenn es dem beherrschenden Unternehmen an der inländischen Betriebsstätte mangelt. Dem steht auch das Urteil des BFH v. 09.02.2011 nicht entgegen, in welchem jegliche Ansässigkeitskriterien des beherrschenden Unternehmens vernachlässigt wurden, da dies anderenfalls gegen das Diskriminierungsverbot des betreffenden DBA (hier: DBA-Großbritannien) verstoßen hätte.[119] Dieses Urteil bezieht sich auf den VZ 1999, in welchem noch nicht das Betriebsstättenerfordernis galt, sondern Sitz und Ort der GL des Organträgers im Inland liegen mussten, daher sind aus dem Urteil keine Auswirkungen auf die Anwendung der heutigen Vorschriften abzuleiten.[120]

Auch unionsrechtlich lässt sich nicht unmittelbar aus dem Urteil des EuGH in der Rs. „SCA Group Holding u.a." oder in der Rs. „Marks & Spencer" ableiten, dass ein Mutterunternehmen ohne inländische Betriebsstätte einbezogen werden muss.[121] In der Rs. „SCA Group Holding u.a." forderte der EuGH lediglich, dass inländische Schwestergesellschaften ihre Gewinne bzw. Verluste miteinander konsolidieren können, auch wenn sie eine auslandsansässige Muttergesellschaft haben, machte jedoch keine Angaben dazu, bei welchem Steuersubjekt die Verrechnung stattzufinden hat.[122] Dennoch wird vereinzelt auf Grundlage dieses Urteils gefordert, dass eine Einkommenszurechnung bei der ausländischen Mutter erfolgt; in diesem Fall müsste der GAV zwischen über- und untergeordnetem Unternehmen abgeschlossen werden.[123] Denkbar wäre in dem Fall auch eine andere Lösung im innerstaatlichen Recht, bei der der Besteuerungszugriff sichergestellt ist, etwa durch Einführung einer beschränkten Steuerpflicht auf das Einkommen der Organgesellschaften.[124] Abkommensrechtlich könnte die Zurechnung als bloßer Austausch des Steuerschuldners interpretiert werden, aufgrund dessen die Besteuerungsrechte nicht abweichend verteilt werden.[125] M.E. ist sich dieser Sichtweise allerdings nicht anzuschließen, da der EuGH diesbezüglich keine Forderungen gestellt hat und Deutschland im Zuge des „Weges des geringsten Widerstands" die bereits von anderen Mitgliedstaaten gewählte Umsetzung über eine Zurechnung bei einer der Schwestergesellschaften wählen sollte, auch wenn diese nicht systemkonform mit der finanziellen Eingliederung ist. Insgesamt ist also m.E. nach aktuellem Rechtsstand kein erfolgsversprechender Weg zu sehen, wonach eine Organschaft mit einem Organträger ohne inländische Betriebsstätte gebildet werden kann.

Horizontale Organschaft

Einen Einfluss auf Gestaltungsmöglichkeiten in dem Fall eines Mutterunternehmens ohne inländische Betriebsstätte kann hingegen das Urteil des EuGH in der Rechtssache „SCA Group Holding u.a." haben. In diesem Urteil entschied der EuGH, dass es einen Verstoß gegen die Niederlassungsfreiheit darstellt, wenn inländische Gesellschaften, die durch eine gemeinsame auslandsansässige Muttergesellschaft gehalten werden, keine steuerliche Ergebniskonsolidierung vornehmen können.[126] Als Folge dieses Urteils haben die Niederlande und Frankreich bereits ihre steuerlichen Vorschriften geändert, um eine horizontale Ergebniskonsolidierung zwischen inländischen Gesellschaften mit einer EU- oder EWR-ansässigen Muttergesellschaft zu ermöglichen.[127] Da Deutschland ebenso wie die Niederlande in dem vor dem EuGH vorliegenden Fall eine inländische Betriebsstätte des Mutterunternehmens fordert, liegt die Frage nach einer entsprechenden Anwendbarkeit des Urteils auf die deutsche Organschaft nahe.[128] Unabhängig von dieser Frage ist in Deutschland eine horizontale Organschaft, also eine Organschaft zwischen Schwestergesellschaften, nach aktuellem Rechtsstand und Rechtsprechung sowohl im Gewerbesteuer- als auch im Körperschaftsteuerrecht nicht möglich, da immer eine (un-)mittelbare Beteiligung in Form eines direkten Über- und Unterordnungsverhältnisses vorliegen muss.[129]

Die Frage, ob das Urteil des EuGH trotz einer gegensätzlichen Entscheidung des BFH v. 22.02.2017 auf die deutsche Organschaft zu übertragen ist, wird im Schrifttum ausführlich diskutiert. Ein Teil der Literatur kritisiert das Urteil des EuGH grundsätzlich, da es den Auslandssachverhalt gegenüber einem rein inländischen Fall bevorzuge und so Inländerdiskriminierung verursache, es insgesamt zu knapp begründet sei, der Gerichtshof evtl. gar nicht zuständig sei, da es sich bei der Betrachtung der beiden inländischen Schwestergesellschaften um keinen grenzüberschreitenden Sachverhalt handele und der EuGH seine Kompetenzen überschritten habe, da es eine Entscheidung des niederländischen Gesetzgebers sei, das System der vertikalen Gewinnkonsolidierung anzuwenden.[130] Kritisiert wird zudem, dass eine Einkommenskonsolidierung zwischen Schwestergesellschaften im Gegen-

[119] Vgl. BFH v. 09.02.2011, I R 54, 55/10, BStBl. II 2012, 106, Rn. 21; Abkommen zwischen der Bundesrepublik Deutschland und dem Vereinigten Königreich Großbritannien und Nordirland zur Vermeidung der Doppelbesteuerung und zur Verhinderung der Steuerverkürzung v. 26.11.1964, BGBl. II 1966, 358, Art. XX Abs. 4, 5.

[120] Vgl. BFH v. 09.02.2011, I R 54, 55/10, BStBl. II 2012, 106; Dötsch/Pung, Organschaft, DB 2013, 307.

[121] Vgl. EuGH v. 02.06.2014, C-39/13, C-40/13, C-41/13, Rs. „SCA Group Holding u.a.", EU:C:2014:1758; Krumm, in: Blümich, KStG, § 14 KStG, Rn. 40 (August 2018); von Brocke/Müller, Holding-Urteil, DStR 2014, 2108; a.A. Herzig/Wagner, Gruppenbesteuerungssysteme, DStR 2006, 11; genauer zur Rs. „SCA Group Holding u.a." s.u. unter Punkt 3.1.2.2.

[122] Vgl. EuGH v. 02.06.2014, C-39/13, C-40/13, C-41/13, Rs. „SCA Group Holding u.a.", EU:C:2014:1758, Rn. 56.

[123] Vgl. Schnitger, Organschaftsreform, IStR 2014, 588 u. 590.

[124] Vgl. Cloer/Kahlenberg, Gruppenbesteuerung, SteuK 2014, 514; Schnitger, Organschaftsreform, IStR 2014, 589.

[125] Vgl. Schnitger, Organschaftsreform, IStR 2014, 588.

[126] Vgl. EuGH v. 2.6.2014, C-39/13, C-40/13, C-41/13, Rs. „SCA Group Holding u.a.", EU:C:2014:1758, Rn. 56.

[127] Vgl. Dölker, in: Dölker/Ergenziger/Feldgen u.a., KStG, § 14 KStG, Rz. 11 (07.02.2020); Walter, Schwestergesellschaften, GmbHR 2015, 182.

[128] Vgl. § 14 Abs. 1 Nr. 2 S. 4 KStG; von Brocke/Müller, Holding-Urteil, DStR 2014, 2107 f.

[129] Vgl. § 14 Abs. 1 S. 1 KStG; Niedersächsisches FG v. 04.09.2007, 6 K 194/07, rkr., EFG 2008, 323, Rz. 12; BFH v. 22.02.2017, I R 35/14, BStBl. II 2018, 33, Rn. 27.

[130] Vgl. Schiefer, C-39/13, C-40/13, C-41/13 (Anm.), EuZW 2014, 703; Rehfeld, Gruppenbesteuerung, IWB 2014, 621 f.; Sydow, Tochtergesellschaften, IStR 2014, 480-482.

satz zur vertikalen Konsolidierung nicht die steuerliche Leistungsfähigkeit des Konzerns abbilde, da die Konsolidierung auf der falschen Ebene stattfinde.[131] Die Gewinnabführung findet in der Regel vertikal von Tochter- zu Mutterunternehmen statt, daher würde es der rechtfertigenden Wirkung der Gewinnabführung für die Organschaft widersprechen, das Einkommen horizontal zuzurechnen.[132] Weiterhin lehnt ein Teil des Schrifttums eine übertragbarkeit auf die deutsche Organschaft ab, da die deutsche Regelung mit dem GAV eine Besonderheit im Vergleich zum niederländischen Recht beinhalte.[133] Außerdem erfolge bei der Organschaft keine Vollkonsolidierung wie in den Niederlanden, sondern eine Einkommenszurechnung, wodurch das BFH-Urteil v. 09.02.2011 zu berücksichtigen sei, da Deutschland einen Verlust des Besteuerungsrechts nicht hinnehmen müsse.[134]

Die überwiegende Meinung im Schrifttum, der sich m.E. anzuschließen ist, besagt jedoch, dass das Urteil auf die deutsche Organschaft anzuwenden sei.[135] Die deutsche Regelung ähnelt sowohl in der Voraussetzung der inländischen Betriebsstätte des Mutterunternehmens als auch im Erfordernis der finanziellen Eingliederung den in Frage stehenden niederländischen Normen.[136] Die finanzielle Eingliederung kann zwischen Schwestergesellschaften typischerweise nicht vorliegen, es soll laut EuGH jedoch ausreichen, dass das Beherrschungsverhältnis zur gemeinsamen Muttergesellschaft vorliegt, welche aber nicht zwingend in die steuerliche Einheit einbezogen wird, d.h. die personelle Einheit von Beherrschung und Konsolidierung wird hier aufgehoben.[137] Folglich müsste wohl weiterhin eine finanzielle Eingliederung zum gemeinsamen Mutterunternehmen vorliegen, lediglich die Einkommenszurechnung würde bei einer der eingegliederten Gesellschaften erfolgen.[138] Ob dies auch ohne eine Gesetzesänderung durch eine teleologische Auslegung möglich ist, ist umstritten.[139] Bei welcher der Gesellschaften die Einkommenszurechnung erfolgt, ist m.E. letztlich zweitrangig, mögliche Anhaltspunkte könnten Größe, Rechtsform oder die finanzielle Fähigkeit zum Ausgleich tatsächlich entstandener Verluste sein.[140]

Fraglich ist allerdings, ob ein GAV zwischen Schwestergesellschaften abgeschlossen werden kann: Grundsätzlich kann der GAV mit einer mittel- oder unmittelbar beteiligten Gesellschaft abgeschlossen werden, in diesem Fall müsste der Abschluss jedoch konzernintern zugunsten Dritter gehen.[141] Die Literatur äußert sich hierzu mit großer Mehrheit nicht ablehnend[142] oder sogar zustimmend[143]. Der GAV zwischen den Schwestergesellschaften ist also mit hoher Wahrscheinlichkeit möglich, wodurch es keinen Grund gibt, dieses Tatbestandsmerkmal bei der horizontalen Organschaft außer Acht zu lassen.[144] Auch gegen die Anwendung der weiteren Voraussetzungen, insbesondere das Erfordernis der Zurechnung zu einem einzigen gewerblichen Unternehmen und das der inländischen Betriebsstätte beim Organträger, sprechen keine Einwände.[145]

Insgesamt ist nach überwiegender Literaturmeinung die Möglichkeit einer Ergebniskonsolidierung zwischen Schwestergesellschaften, wenn das Mutterunternehmen keine inländische Betriebsstätte hat, aus unionsrechtlicher Sicht geboten.[146] Bei einem EU-/EWR-ansässigen Mutterunternehmen mit einer deutschen Tochter- und zwei deutschen Enkelgesellschaften ist die Anwendung des Urteils nicht erforderlich, da in diesem Fall die Konsolidierung über die deutsche Tochtergesellschaft erfolgen kann und keine horizontale Organschaft erforderlich ist. [147]

Hierbei ist zu beachten, dass der EuGH nicht ausdrücklich eine horizontale Einkommensverrechnung der Schwestergesellschaften untereinander verlangt hat, sondern lediglich gefordert hat, dass eine Verrechnung in irgendeiner Form stattfinden muss, wobei er offen gelassen hat, ob die Verrechnung bei der Schwestergesellschaft oder der Mutterge-

[131] Vgl. Sydow, Tochtergesellschaften, IStR 2014, 483.

[132] Vgl. van Lishaut, Reform, FR 2009, 1031.

[133] Vgl. Möhlenbrock, Niederlassungsfreiheit, DB 2014, 1583; Schiefer, C-39/13, C-40/13, C-41/13 (Anm.), EuZW 2014, 703.

[134] Vgl. Sydow, Tochtergesellschaften, IStR 2014, 484 f.; Schober, in: Musil/Weber-Grellet, Steuerrecht, § 14 KStG, Rn. 123 (2019).

[135] Vgl. von Brocke/Müller, Holding-Urteil, DStR 2014, 2109; Cloer/Kahlenberg, Gruppenbesteuerung, SteuK 2014, 513 f.; Lieber, C-39/13, C-40/13, C-41/13 (Anm.), ISR 2014, 289; Schnitger, Organschaftsreform, IStR 2014, 587 f.; Walter, Schwestergesellschaften, GmbHR 2015, 183 f.; Hackemann, Taxation, in: Richelle/Schön/Traversa, Union, 2013, 128.

[136] Vgl. § 14 Abs. 1 S. 1 Nr. 1, S. 1 Nr. 2 S. 4 KStG; EuGH v. 02.06.2014, C-39/13, C-40/13, C-41/13, Rs. „SCA Group Holding u.a.“, EU:C:2014:1758 Rn. 4.

[137] Vgl. EuGH v. 02.06.2014, C-39/13, C-40/13, C-41/13, Rs. „SCA Group Holding u.a.“, EU:C:2014:1758; EuGH, Schlussanträge der Generalanwältin Kokott v. 27.02.2014, C-39/13, Rs. „SCA Group Holding u.a.“, EU:C:2014:104, Tz. 76-78; von Brocke/Müller, Holding-Urteil, DStR 2014, 2109.

[138] Vgl. von Brocke/Müller, Holding-Urteil, DStR 2014, 2109; Walter, Schwestergesellschaften, GmbHR 2015, 183.

[139] Befürwortend vgl. Walter, Schwestergesellschaften, GmbHR 2015, 183; ablehnend vgl. Gosch, Gewinngemeinschaft, IStR 2018, 749.

[140] Vgl. Walter, Schwestergesellschaften, GmbHR 2015, 183.

[141] Vgl. von Brocke/Müller, Holding-Urteil, DStR 2014, 2109.

[142] Vgl. Deilmann, in: Hölters, AktG, § 291 AktG, Rn. 11 (2017); Koch, in: Hüffer/Koch, AktG, § 291 AktG, Rn. 25 (2020); a.A. Möhlenbrock, Niederlassungsfreiheit, DB 2014, 1583.

[143] Vgl. OLG Nürnberg v. 17.01.1996, 12 U 2801/95, AG 1996, 228; OLG Frankfurt/M. v. 26.08.2009, 23 U 69/08, AG 2010, 368, Rn. 91; Altmeppen, in: Goette/Habersack, AktG, § 291 AktG, Rn. 18 (2015); Veil, in: Spindler/Stilz, AktG, § 291 AktG, Rn. 6 (2019); Emmerich, in: Emmerich/Habersack, Konzernrecht, § 291 AktG, Rn. 10 (2019); Paschos, in: Henssler/Strohn, Gesellschaftsrecht, § 291 AktG, Rn. 6 (2019); Koch, in: Hüffer/Koch, AktG, § 291 AktG, Rn. 25 (2020); Walter, in: Bott/Walter, KStG, § 14 KStG, Rz. 580 (Februar 2020); Thoma, Unternehmensverträge, in: Krieger/Lutter/Schmidt, FS Hoffmann-Becking, 2013, 1240 f.; Walter, Gewinnabführungsvertrag, DB 2014, 2016 f.; Walter, Schwestergesellschaften, GmbHR 2015, 182; a.A. van Venrooy, Unternehmensverträge, BB 1986, 615.

[144] Vgl. Walter, Schwestergesellschaften, GmbHR 2015, 183.

[145] Vgl. Walter, Schwestergesellschaften, GmbHR 2015, 183.

[146] Vgl. von Brocke/Müller, Holding-Urteil, DStR 2014, 2110; Cloer/Kahlenberg, Gruppenbesteuerung, SteuK 2014, 514; Lieber, C-39/13, C-40/13, C-41/13 (Anm.), ISR 2014, 280; Schnitger, Organschaftsreform, IStR 2014, 587; Juretzek, C-41/13 (Anm.), GWR 2014, 336; a.A. Möhlenbrock, Niederlassungsfreiheit, DB 2014, 1583; Sydow, Tochtergesellschaften, IStR 2014, 484 f.; Walter, Schwestergesellschaften, GmbHR 2015, 184; Kessler/Arnold, Organschaft, IStR 2016, 228; zur Pflicht zur Zurechnung bei der Muttergesellschaft s.o. unter Punkt 3.1.2.1.

[147] Vgl. Walter, Schwestergesellschaften, GmbHR 2015, 184.

sellschaft stattfindet.[148] Lediglich in der Praxis führt dies in der einfachsten Gestaltung zu einer horizontalen Organschaft, damit der betreffende Staat durch die Zurechnung zum auslandsansässigen Mutterunternehmen nicht das Besteuerungsrecht verliert.[149]

Der BFH hat hingegen, ohne genauer auf die Gründe für diese Entscheidung oder den Bezug zum Urteil des EuGH in der Rs. „SCA Group Holding u.a." einzugehen, entschieden, dass keine Diskriminierung darin zu erkennen sei, dass eine Organschaft zwischen Schwestergesellschaften nicht zugelassen wird.[150] Die Tatsache, dass der BFH hierzu kein Vorabentscheidungsverfahren beim EuGH eingeleitet hat, könnte m.E. einen Verstoß gegen Art. 101 Abs. 1 S. 2 GG und Art. 267 Abs. 3 AEUV darstellen.[151] Es ist stark anzuzweifeln, ob die Sichtweise des BFH auch vom EuGH anerkannt würde; es bleibt für den Steuerpflichtigen also nur zu hoffen, dass diese Angelegenheit vor dem EuGH verhandelt werden wird.[152]

3.1.3. Auslandssachverhalt mit Inlandsbezug

Obwohl die Organschaft grds. grenzüberschreitende Sachverhalte eher vernachlässigt, ist die Bildung einer Organschaft eines ausländischen Mutterunternehmens mit einer ausländischen Tochtergesellschaft nicht von vorherein auszuschließen. Hierfür müssen jedoch zwei Voraussetzungen vorliegen: Zunächst muss das Mutterunternehmen eine inländische Betriebsstätte besitzen, welcher die Beteiligung an der Tochtergesellschaft zuzuordnen ist, denn ansonsten würde es dem Sachverhalt ohnehin an einem Inlandsbezug mangeln.[153] Zudem muss die ausländische Tochtergesellschaft ihren Ort der GL im Inland haben, um als Organgesellschaft qualifiziert zu sein.[154] Liegt seitens der Tochtergesellschaft kein Ort der GL im Inland vor, ist m.E. das unten dargestellte Urteil des EuGH zu der Rs. „Marks & Spencer" zu finalen Verlusten mitsamt aller Voraussetzungen auf den Sachverhalt anzuwenden, sodass in diesem Fall nur die finalen Verluste der Tochtergesellschaft von den Gewinnen des Mutterunternehmens abziehbar sind.[155]

3.2. Drittland

Für ein Mutterunternehmen in einem Drittland ist die Rechtslage nicht allzu sehr von der Rechtslage im Fall eines EU- oder EWR-ansässigen Mutterunternehmens zu unterscheiden. Seit der Gesetzesänderung 2013 ist die Ansässigkeit des Organträgers nicht mehr relevant, solange er eine inländische Betriebsstätte hat, der die Beteiligung an der Organgesellschaft zuzuordnen ist.[156] Der Fall eines Mutterunternehmens mit einer inländischen Betriebsstätte, welcher die Beteiligung an der Tochtergesellschaft zuzuordnen ist, ist also bei einer Ansässigkeit des Mutterunternehmens in einem Drittland in keiner Weise anders zu behandeln als bei einer Ansässigkeit in einem EU- oder EWR-Staat. Hierbei sind die gleichen kritischen Punkte zu beachten, die bereits oben dargelegt wurden, insbesondere die Zuordnung zur inländischen Betriebsstätte zur Sicherung des inländischen Steuersubstrats und der Abschluss eines GAV.[157]

Bei einem Mutterunternehmen aus einem Drittland ohne inländische Betriebsstätte ist jedoch, anders als bei einem Mutterunternehmen aus dem Gemeinschaftsgebiet oder dem Gebiet des EWR, die Bildung einer horizontalen Organschaft nicht zu erwägen, da diese wie oben ausgeführt auf einem Urteil des EuGH basiert, welches für Gesellschaften aus Staaten außerhalb der EU bzw. des EWR keine Bindungswirkung entfaltet.[158] Auch ist hier ein Auslandssachverhalt mit Inlandsbezug aufgrund der nicht erfüllten Ansässigkeitskriterien der Organgesellschaft ausgeschlossen, diese dürfte ihren Sitz also nur in einem EU- oder EWR-Staat und den Ort der GL im Inland haben, womit es ihr an jeglichem Drittlandsbezug fehlen würde.[159]

4. Outbound-Sachverhalt

4.1. EU-/EWR-Staat

4.1.1. Auslandsansässige Organgesellschaft

Rechtsstand

In einem Outbound-Sachverhalt mit einem inlandsansässigen Mutterunternehmen und einer auslandsansässigen Tochtergesellschaft ohne Inlandsbezug kann grds. aufgrund der nicht erfüllten Voraussetzung des inländischen Ortes der GL bei der Tochtergesellschaft keine Organschaft gebildet werden.[160] Auch unionsrechtlich besteht kein Zwang zur Berücksichtigung laufender Verluste einer ausländischen Tochtergesellschaft bei der inländischen Besteuerung.[161] Hierbei bildet jedoch die Rechtsprechung des EuGH zu finalen Verlusten eine Ausnahme.[162]

Der EuGH hat in der Rs. „Marks & Spencer" in einem viel beachteten Urteil entschieden, dass es zwar mit der in

[148] Vgl. EuGH v. 02.06.2014, C-39/13, C-40/13, C-41/13, Rs. „SCA Group Holding u.a.", EU:C:2014:1758.

[149] Vgl. von Brocke/Müller, Holding-Urteil, DStR 2014, 2109; Kessler/Arnold, Organschaft, IStR 2016, 228.

[150] Vgl. BFH v. 22.02.2017, I R 35/14, BStBl. II 2018, 33, Rn. 27; kritisch Gosch, Gewinngemeinschaft, IStR 2018, 740.

[151] Vgl. BVerfG v. 06.10.2017, 2 BvR 987/16, NJW 2018, 606, Rn. 9.

[152] Vgl. Walter, Gewinngemeinschaft, Der Konzern 2017, 334.

[153] Vgl. § 14 Abs. 1 S. 1 Nr. 2 S. 4 KStG; zum Betriebsstättenerfordernis s.o. unter Punkt 3.1.1.

[154] Vgl. § 14 Abs. 1 S. 1 KStG.

[155] Vgl. EuGH v. 31.02.2005, C-446/03, Rs. „Marks & Spencer", EU:C:2005:763; genauer zu dem Urteil s.u. unter Punkt 4.1.1.

[156] Vgl. Gesetz zur änderung und Vereinfachung der Unternehmensbesteuerung und des steuerlichen Reisekostenrechts v. 20.02.2013, BGBl. I 2013, 285; Breuninger, Grundlagen, in: Prinz/Witt, Organschaft, 2. Aufl. 2019, 911, Rz. 25.2; Prinz/Ludwig, Gewinnabführungsvertrag, DB 2020, 1022.

[157] Vgl. Breuninger, Grundlagen, in: Prinz/Witt, Organschaft, 2. Aufl., Rz. 25.3 u. 25.4; s.o. unter Punkt 3.1.1.

[158] Vgl. Walter, Schwestergesellschaften, GmbHR 2015, 183.

[159] Genauer zum Auslandssachverhalt mit Inlandsbezug vgl. oben unter Punkt 3.1.3; zu einer im Drittland ansässigen Tochtergesellschaft vgl. unten unter Punkt 4.2

[160] Vgl. § 14 Abs. 1 S. 1 KStG; Balmes/Brück/Ribbrock, Organschaftsbesteuerung, BB 2005, 969.

[161] Vgl. EuGH v. 25.02.2010, C-337/08, Rs. „X Holding", EU:C:2010:89, Rn. 40; Frotscher, Organschaft, IStR 2011, 702.

[162] Genauer zum Begriff der finalen Verluste vgl. unten unter Punkt 4.1.1.3.

Art. 49 u. 54 AEUV (ehem. Art. 43 u. 48 EG) kodifizierten Niederlassungsfreiheit vereinbar ist, wenn ein Staat den Abzug der Verluste einer ausländischen Tochtergesellschaft vom inländischen Gewinn der Muttergesellschaft nicht zulässt.[163] Es stellt jedoch einen Verstoß gegen die Niederlassungsfreiheit dar, wenn eine gebietsfremde Tochtergesellschaft einer gebietsansässigen Muttergesellschaft die in ihrem Sitzstaat vorgesehenen Möglichkeiten für die Verlustnutzung im betreffenden und in vorherigen Steuerzeiträumen ausgeschöpft hat und es keine Möglichkeit gibt, dass die Verluste im Sitzstaat von der Tochtergesellschaft selbst oder von einem Dritten genutzt werden können.[164] Der Anwendungsbereich dieser Behandlung von finalen Verlusten wurde mit einem späteren Urteil auch auf Betriebsstätten im Gemeinschaftsgebiet ausgedehnt.[165]

In Deutschland haben sich bereits mehrmals die Finanzgerichte mit der Anwendung des Urteils „Marks & Spencer“ auf die Organschaft befasst, bisher wurde jedoch kein höchstrichterliches Urteil getroffen.[166] Zunächst einmal ist festzuhalten, dass nach h.M. die Grundsätze des o.g. Urteils „Marks & Spencer“ des EuGH grds. auf die deutsche Organschaft anzuwenden sind und sich auch die Finanzgerichte in dieser Hinsicht positiv geäußert haben, ohne jedoch eine endgültige Entscheidung zu treffen.[167] Die deutsche Organschaft verletzt also ihrem Gesetzestext nach die unionsrechtliche Niederlassungsfreiheit ohne Rechtfertigung.[168] Nach Meinung der Finanzgerichte ist in diesem Kontext auf das Erfordernis des in den jeweiligen Streitjahren noch geltenden doppelten Inlandsbezuges der Organgesellschaft, d.h. das Erfordernis von Sitz und Ort der GL im Inland, zu verzichten.[169] Der Verzicht auf dieses Erfordernis ist m.E. unzweifelhaft auch auf den heutigen einfachen Inlandsbezug, d.h. das Erfordernis des Ortes der GL im Inland, anzuwenden, da das Urteil „Marks & Spencer“ sich auf eine auslandsansässige Tochtergesellschaft ohne jegliche Ansässigkeitsmerkmale im Inland bezieht und die betreffende britische Regelung Sitz oder wirtschaftliche Betätigung im Vereinigten Königreich forderte.[170]

Die Ansässigkeitsmerkmale der ausländischen Tochtergesellschaft sollten also kein Hindernis bei der Verrechnung von finalen Verlusten darstellen, ebenso ist davon auszugehen, dass die finanzielle Eingliederung und die notwendigen Voraussetzungen zur Bildung einer Organschaft bei dem Mutterunternehmen vorliegen, welche auch bei der Verrechnung von finalen Verlusten befolgt werden müssen.[171] Lediglich bei einer mittelbaren Beteiligung an der Tochter- oder Enkelgesellschaft, bei der die Verluste entstanden sind, sind Ansässigkeitsmerkmale relevant: Hat die vermittelnde Gesellschaft bzw. haben die vermittelnden Gesellschaften ihren Sitz im Ansässigkeitsstaat der Gesellschaft, bei der die Verluste entstanden sind, gilt die Rechtsprechung des EuGH in der Rs. „Marks & Spencer“ entsprechend, eine Verrechnung der finalen Verluste ist geboten.[172] Liegt der Sitz der vermittelnde(n) Gesellschaft(en) in einem anderen Staat als dem Ansässigkeitsstaat der Tochter- oder Enkelgesellschaft, ist unionsrechtlich jedoch keine Verrechnung geboten, da der Konzern in diesem Fall ggf. wählen könnte, in welchem Staat die Verluste verrechnet werden.[173]

Umstritten ist, ob die Rechtsprechung zu finalen Verlusten auch auf die Gewerbesteuer anzuwenden ist.[174] Hiergegen spricht die territoriale Begrenzung des GewStG auf das Inland.[175] Zumindest für finale Betriebsstättenverluste wurde eine Abzugspflicht in der Gewerbesteuer vom BFH bejaht, auch wenn die Vorinstanzen teilweise anders entschieden hatten.[176] Das Schrifttum hat sich mit dieser Fragestellung bisher nicht ausführlich auseinandergesetzt, m.E. sind jedoch die im Urteil des Schleswig-Holsteinischen FG v. 13.03.2019

[163] Vgl. Art. 43 u. 48 EG; Art. 49 u. 54 AEUV; EuGH v. 31.02.2005, C-446/03, Rs. „Marks & Spencer“, EU:C:2005:763, Rn. 59.

[164] Vgl. EuGH v. 31.02.2005, C-446/03, Rs. „Marks & Spencer“, EU:C:2005:763, Rn. 59.

[165] Vgl. EuGH v. 15.05.2008, C-414/06, Rs. „Lidl Belgium“, EU:C:2008:278, Rn. 54; Dölker, in: Dölker/Ergenziger/Feldgen u.a., KStG, § 14 KStG, Rz. 7 (07.02.2020).

[166] Vgl. Niedersächsisches FG v. 11.02.2010, 6 K 406/08, rkr., EFG 2010, 815; FG Rheinland-Pfalz v. 17.03.2010, 1 K 2406/07, rkr., EFG 2010, 1632; Schleswig-Holsteinisches FG v. 13.03.2019, 1 K 218/15, Rev. eingelegt (Az. des BFH: I R 26/19), EFG 2019, 1466.

[167] Vgl. Niedersächsisches FG v. 11.02.2010, 6 K 406/08, rkr., EFG 2010, 815, Rn. 38 u. 41; FG Rheinland-Pfalz v. 17.03.2010, 1 K 2406/07, rkr., EFG 2010, 1632, Rn. 47; Schleswig-Holsteinisches FG v. 13.03.2019, 1 K 218/15, Rev. eingelegt (Az. des BFH: I R 26/19), EFG 2019, 1466, Rn. 59-61; gl.A. Herzig/Wagner, Gruppenbesteuerungssysteme, DStR 2006, 9; Hey, Organschaft, GmbHR 2006, 118; Englisch, C-446/03 (Anm.), IStR 2006, 23; a.A. Thiel, Ertragsbesteuerung, DB 2004, 2605.

[168] Vgl. Behrens, Anpassung, Börsen-Zeitung Nr. 244, 6; Balmes/Grammel/Sedemund, Betriebsstättenverluste, BB 2006, 1477 f.; Herzig/Wagner, Gruppenbesteuerungssysteme, DStR 2006, 9; Pache/Englert, Konzernbesteuerungsrecht, IStR 2007, 848 f.; Wittkowski, Verlustverrechnung, 2008, 77; von Brocke, FG-Entscheidungen, DStR 2010, 964; Dötsch/Pung, Organschaft, DB 2013, 306; a.A. Thiel, Ertragsbesteuerung, DB 2004, 2605; zweifelnd Balmes/Brück/Ribbrock, Organschaftsbesteuerung, BB 2005, 969.

[169] Vgl. Niedersächsisches FG v. 11.02.2010, 6 K 406/08, rkr., EFG 2010, 815, Rn. 46; FG Rheinland-Pfalz v. 17.03.2010, 1 K 2406/07, rkr., EFG 2010, 1632, Rn. 76; Schleswig-Holsteinisches FG v. 13.03.2019, 1 K 218/15, Rev. eingelegt (Az. des BFH: I R 26/19), EFG 2019, 1466, Rn. 71; gl.A. Kleinert/Nagler/Rehm, Gewinnbesteuerung, DB 2005, 1872; Kleinert/Nagler, VI R 32/04 (Anm.), DB 2005, 2792; Balmes/Brück/Ribbrock, Organschaftsbesteuerung, BB 2005, 968; Herzig/Wagner, Gruppenbesteuerungssysteme, DStR 2006, 9; Altrichter-Herzberg/Nuernberger, Verlustnutzung, GmbHR 2006, 466; a.A. Mitschke, Ergebnisabführungsvertrag, IStR 2011, 188.

[170] Vgl. EuGH v. 31.02.2005, C-446/03, Rs. „Marks & Spencer“, EU:C:2005:763, Rn 22 f.; EuGH, Schlussanträge des Generalanwaltes Maduro v. 07.04.2005, C-446/03, Rs. „Marks & Spencer“, EU:C:2005:201, Rn. 9.

[171] Vgl. Niedersächsisches FG v. 11. 2. 2010, 6 K 406/08, rkr., EFG 2010, 815, Rn. 58.

[172] Vgl. EuGH v. 19.06.2019, C-608/17, Rs. „Holmen“, EU:C:2019:511, Rn. 33.

[173] Vgl. EuGH v. 19.06.2019, C-608/17, Rs. „Holmen“, EU:C:2019:511, Rn. 26-29.

[174] Vgl. Henkel, Beteiligung, in: Mössner/Baumhoff/Greif, Steuerrecht, 5. Aufl. 2018, Rz. 7.105.

[175] Vgl. van Lishaut, Reform, FR 2009, 1036; Heckerodt, Betriebsstättenverluste, IStR 2019, 175.

[176] Vgl. BFH v. 09.06.2010, I R 100/09, BStBl. II 2010, 1065, Rn. 26; FG Hamburg v. 18.11.2009, 6 K 147/08, Rev. eingelegt (Az. des BFH: I R 107/09), EFG 2010, 265, Rn. 31. Die Revision wurde in diesem Punkt als unbegründet zurückgewiesen, vgl. BFH v. 09.06.2010, I R 100/09, BStBl. II 2010, 1065; FG Hessen v. 04.09.2018, 4 K 385/17, Rev. eingelegt (Az. des BFH: I R 32/18), EFG 2018, 1876, Rn. 30.

genannten Grundsätze auf die Gewerbesteuer zu übertragen, da die Klägerin die finalen Verluste auch für die Gewerbesteuer abziehen wollte und das FG auf diese Forderung nicht abweichend eingegangen ist.[177] Demnach sind finale Verluste sowohl bei der Körperschaft- als auch bei der Gewerbesteuer abziehbar, wenn die unten ausgeführten Voraussetzungen vorliegen.[178]

Gewinnabführungsvertrag

Fraglich ist, wie mit dem GAV umzugehen ist, da es in der Regel nicht möglich ist, einen solchen mit einer ausländischen Tochtergesellschaft abzuschließen, da das ausländische Recht eine so weitreichende Aufgabe der Autonomie häufig nicht akzeptiert.[179] Dies war auch in den Urteilen der Finanzgerichte der vorherrschende Aspekt, mit welchem sich die Gerichte beschäftigten.[180] Hierzu ist die Vergleichspaarbildung zur überprüfung einer gemeinschaftsrechtswidrigen Ungleichbehandlung entscheidend: Nach der überprüfung, ob eine nationale Regelung grundsätzlich die Ausübung der Grundfreiheiten nach dem AEUV einschränkt, ist zu prüfen, ob eine mit einem Inlandssachverhalt vergleichbare Lage vorliegt sowie ob es Rechtfertigungsgründe für diese Ungleichbehandlung gibt.[181] Bildet man das Vergleichspaar zwischen einer Organschaft zwischen einem inländischen Mutter- und einem ausländischen Tochterunternehmen ohne GAV und einem inländischen Mutter- und einem inländischen Tochterunternehmen ohne GAV, so läge keine Ungleichbehandlung vor, da in beiden Fällen die Rechtsfolge, nämlich keine Anerkennung der Organschaft, die gleiche ist.[182] Demnach könnte man allerdings durch die Schaffung eines spezifischen inländischen Instituts beinahe immer eine Vergleichbarkeit zwischen Inlands- und Auslandssachverhalt ausschließen, wodurch die Grundfreiheiten ihres Zwecks entleert würden.[183] Die andere, m.E. zutreffendere Möglichkeit ist, das Vergleichspaar zwischen einem inländischen Mutter- und einem ausländischen Tochterunternehmen ohne GAV und einem inländischen Mutter- und einem inländischen Tochterunternehmen mit GAV zu bilden, da der GAV als verdeckte Diskriminierung nicht selbst Teil der Vergleichbarkeitsprüfung sein kann.[184]

Die Finanzgerichte haben in dieser Frage nicht entschieden, da sie übereinstimmend der Meinung waren, dass im Fall einer Ungleichbehandlung, wie sie beim zweiten o.g. Vergleichspaar vorläge, das Erfordernis des GAV aufgrund seiner essentiellen Bedeutung für die Organschaft nur geltungserhaltend reduziert und nicht vollständig aufgegeben werden müsse.[185] Da in den vorliegenden Sachverhalten jeweils kein GAV in vollständiger oder reduzierter Form abgeschlossen worden war, konnten die Gerichte diese Frage offen lassen.[186]

Ein reduzierter GAV soll nach Meinung der Gerichte eine schuldrechtliche Verpflichtung zur Verlustübernahme seitens des Mutterunternehmens sein, welche nach h.M. auch über die Grenze hinweg möglich ist.[187] Diese Verpflichtung müsste, ebenso wie der GAV, über mindestens 5 Jahre abgeschlossen werden.[188] Eine freiwillige Verlustübernahme ohne rechtliche Verpflichtung soll nicht ausreichen, da etwa im Inlandsfall auch keine Verpflichtung ausreicht, die sich aus zivilrechtlicher Rechtsprechung ergibt und folglich eine solch weitreichende Erleichterung im Auslandssachverhalt unverhältnismäßig wäre.[189]

Diese Meinung wurde kritisiert, da einige Stimmen im Schrifttum entsprechend des Normzwecks ein Abstellen auf das tatsächliche Verhalten des Mutterunternehmens oder eine Optionsmöglichkeit für den Steuerpflichtigen forderten,[190] da die Forderung einer schuldrechtlichen Verlustüber-

[177] Vgl. Schleswig-Holsteinisches FG v. 13.03.2019, 1 K 218/15, Rev. eingelegt (Az. des BFH: I R 26/19), EFG 2019, 1466, Rn. 1.

[178] Vgl. unter Punkt 4.1.1.2 und 4.1.1.3.

[179] Vgl. Schleswig-Holsteinisches FG v. 13.03.2019, 1 K 218/15, Rev. eingelegt (Az. des BFH: I R 26/19), EFG 2019, 1466, Rn. 22, 73 f.; Herzig/Wagner, Zukunft, DB 2005, 5; C. Witt, Konzernbesteuerung, FR 2009, 1047; Hoene, Gewinnabführungsvertrag, IStR 2012, 463; Kessler/Arnold, Organschaft, IStR 2016, 231.

[180] Vgl. Niedersächsisches FG v. 11.02.2010, 6 K 406/08, rkr., EFG 2010, 815; FG Rheinland-Pfalz v. 17.03.2010, 1 K 2406/07, rkr., EFG 2010, 1632; Schleswig-Holsteinisches FG v. 13.03.2019, 1 K 218/15, Rev. eingelegt (Az. des BFH: I R 26/19), EFG 2019, 1466.

[181] Vgl. etwa EuGH v. 31.02.2005, C-446/03, Rs. „Marks & Spencer", EU:C:2005:763, Rn. 34 f.; Kokott, Steuerrecht, 2018, 131-133; Maack/Kersten, GAV, DStR 2019, 2282.

[182] Vgl. FG Rheinland-Pfalz v. 17.03.2010, 1 K 2406/07, rkr., EFG 2010, 1632, Rn. 67 f.; Schleswig-Holsteinisches FG v. 13.03.2019, 1 K 218/15, Rev. eingelegt (Az. des BFH: I R 26/19), EFG 2019, 1466, Rn. 65-68; gl.A. Mitschke, Nulltarif, DStR 2010, 1369 f.

[183] Vgl. Krumm, in: Blümich, KStG, § 14 KStG, Rn. 39 (August 2018).

[184] Vgl. Glahe, Organschaft, IStR 2012, 130; Graw, Verluste, DB 2010, 2471; Micker, Inlandsbezug, IWB 2013, 312; Maack/Kersten, GAV, DStR 2019, 2283.

[185] Vgl. FG Rheinland-Pfalz v. 17.03.2010, 1 K 2406/07, rkr., EFG 2010, 1632, Rn. 84; Schleswig-Holsteinisches FG v. 13.03.2019, 1 K 218/15, Rev. eingelegt (Az. des BFH: I R 26/19), EFG 2019, 1466, Rn. 69.

[186] Vgl. FG Rheinland-Pfalz v. 17.03.2010, 1 K 2406/07, rkr., EFG 2010, 1632, Rn. 84, 86; Schleswig-Holsteinisches FG v. 13.03.2019, 1 K 218/15, Rev. eingelegt (Az. des BFH: I R 26/19), EFG 2019, 1466, Rn. 69.

[187] Vgl. Niedersächsisches FG v. 11.2.2010, 6 K 406/08, rkr., EFG 2010, 815, Rn. 52, 54-57; FG Rheinland-Pfalz v. 17.3.2010, 1 K 2406/07, rkr., EFG 2010, 1632, Rn. 85; Schleswig-Holsteinisches FG v. 13.03.2019, 1 K 218/15, Rev. eingelegt (Az. des BFH: I R 26/19), EFG 2019, 1466, Rn. 76, 82-84.; zustimmend Kolbe in H/H/R, KStG, § 14 KStG, Anm. 12 (Januar 2015); Balmes/Brück/Ribbrock, Verlustnutzung, BB 2006, 188 f.; Englisch, C-446/03 (Anm.), IStR 2006, 23; Mitschke, Nulltarif, DStR 2010, 1371; Prinz/Ludwig, Gewinnabführungsvertrag, DB 2020, 1027 f.; zum reduzierten GAV vgl. Rödder/Liekenbrock, in: Rödder/Herlinghaus/Neumann, KStG, § 14 KStG, Rz. 89 (2015); Krumm, in: Blümich, KStG, § 14 KStG Rz. 39 (August 2018); Micker, Verlustverrechnung, IWB 2013, 313 f.; Schnitger, Betriebsausgaben, IStR 2015, 773; Kessler/Arnold, Organschaft, IStR 2016, 231; Stöber, Verlustverrechnung, DStZ 2016, 589;.

[188] Vgl. Niedersächsisches FG v. 11. 2. 2010, 6 K 406/08, rkr., EFG 2010, 815, Rn. 58.

[189] Vgl. BFH v. 22.02.2006, I R 74/05, BFH/NV 2006, 1513; Niedersächsisches FG v. 11. 2. 2010, 6 K 406/08, rkr., EFG 2010, 815, Rn. 56; FG Rheinland-Pfalz v. 17.03.2010, 1 K 2406/07, rkr., EFG 2010, 1632, Rn. 88; Schleswig-Holsteinisches FG v. 13.03.2019, 1 K 218/15, Rev. eingelegt (Az. des BFH: I R 26/19), EFG 2019, 1466, Rn. 80; gl. A. Altrichter-Herzberg/Nuernberger, Verlustnutzung, GmbHR 2006, 467; Prinz/Ludwig, Gewinnabführungsvertrag, DB 2020, 1026 f.

[190] Zum freiwilligen Ausgleich vgl. Raupach/Pohl, Konzernsteuerrecht, NZG 2005, 492; von Brocke, FG-Entscheidungen, DStR 2010, 966; Maack/Kersten, GAV, DStR 2019, 2283; zur Optionsmöglichkeit vgl. Altrichter-Herzberg/Nuernberger, Verlustnutzung, GmbHR 2006, 467 f.

nahmeverpflichtung kein verminderter organisationsrechtlicher GAV sei, sondern eine neue Voraussetzung darstelle.[191] Die Tatsache, dass dadurch die Voraussetzungen bei einem Auslandssachverhalt verglichen mit denen eines Inlandssachverhaltes deutlich weniger hoch gesetzt wären, stelle keine Inländerdiskriminierung dar, da solche Unterschiede bei einem nicht vollständig harmonisierten Binnenmarkt zu akzeptieren seien.[192] Zu einem ähnlichen Ergebnis kommt die Sichtweise auf den GAV nicht als eigene Voraussetzung für die Organschaft, sondern als typisierendes Merkmal, um eine wirtschaftliche Einheit zwischen Mutter- und Tochtergesellschaft auch ohne Prüfung des Einzelfalls festzustellen, weshalb bei finalen Verlusten ein Motivtest durchzuführen sei.[193] Diese Sichtweise überzeugt m.E. nicht, da es auch bei einem grenzüberschreitenden Sachverhalt möglich sein sollte, mit einer Verlustübernahmeverpflichtung ein solches typisierendes Merkmal zu fordern, um auch im Sinne der Rechtssicherheit für den Steuerpflichtigen eine Einzelfallprüfung zu vermeiden. Ein anderer Kritikpunkt ist, dass der verminderte GAV zu einer zu großen Rechtsunsicherheit führe und es nicht klar sei, ob eine Verlustübernahmeverpflichtung nach § 301 AktG in allen EU- bzw. EWR-Staaten abgeschlossen werden könne.[194]

Andererseits wurde allerdings auch Kritik geäußert, wonach das GAV-äquivalent eine zu weitgehende Erleichterung gegenüber dem GAV-Erfordernis im Inlandsfall sei.[195] Dieses Argument wird jedoch m.E. zutreffend dadurch entkräftet, dass auch die durch das GAV-äquivalent hervorgerufenen Rechtsfolgen weniger weitreichende sind, als es bei einer Organschaft im Inland der Fall ist, denn es können lediglich finale Verluste, nicht auch laufende, abgezogen werden.[196] Zudem wurde vorgebracht, eine reine Verlustübernahmeverpflichtung werde der spiegelbildlichen Betrachtung von Jahresüberschuss und Jahresfehlbetrag nicht gerecht.[197]

Die Verlustübernahmeverpflichtung kann dazu führen, dass im Sitzstaat der Tochtergesellschaft keine finalen Verluste entstehen, etwa weil die Ausgleichszahlungen als Einnahme eingestuft werden.[198] Dies wird insofern kritisiert, als dass dem Steuerpflichtigen dadurch der Rechtsweg verwehrt bleibt, da in diesem Fall eine Abzugsfähigkeit im Inland mangels finaler Verluste nicht in Frage kommt.[199]

Begriff und Abziehbarkeit finaler Verluste

Weiterhin verbleibt die Frage, wann genau finale Verluste vorliegen. Grundsätzlich muss die Verlustnutzung im Ansässigkeitsstaat der Tochtergesellschaften aus tatsächlichen Gründen ausgeschlossen sein.[200] Besteht also noch eine rechtliche Möglichkeit, die Verluste zu nutzen, die aber aus tatsächlichen Gründen nicht ausgenutzt werden kann, können dennoch finale Verluste vorliegen.[201]

Wie im Sachverhalt, der dem Urteil „Marks & Spencer“ zugrunde lag, kann zunächst einmal eine Liquidation der Tochtergesellschaft zur Entstehung von finalen Verlusten führen, wenn zuvor alle Möglichkeiten zum Verlustabzug ausgenutzt wurden und durch die Liquidation ein Abzug der verbleibenden Verluste unmöglich wird.[202] Folgt der Staat der Tochtergesellschaft der Sitztheorie, kann schon eine grenzüberschreitende Verlegung des Verwaltungssitzes zu einer Zwangsliquidation und somit der Entstehung von finalen Verlusten im ursprünglichen Ansässigkeitsstaat der Tochtergesellschaft führen.[203] Die unternehmerische Tätigkeit im Ausland muss aber nicht zwangsläufig aufgegeben werden, etwa durch Verschmelzen der Tochtergesellschaft auf die Muttergesellschaft oder durch Umwandlung der Tochtergesellschaft in eine PersG, sofern bei diesen Vorgängen der Verlustvortrag nicht auf die Mutter- bzw. die neue PersG oder ihre Gesellschafter übergeht.[204] Hierbei ist jedoch immer darauf zu achten, dass kein Gestaltungsmissbrauch nach § 42 AO entsteht, was der Fall wäre, wenn verständige Personen bei dem gegebenen wirtschaftlichen Sachverhalt und der Zielsetzung andere Handlungsmöglichkeiten ergreifen würden.[205] Unklar ist, wie der Steuerpflichtige genau nachweisen soll, dass alle Verlustverrechnungsmöglichkeiten ausgeschöpft wurden, da etwa bei einem Verkauf das Mutterunternehmen diese Beweise nicht mehr erbringen kann.[206]

[191] Vgl. von Brocke/Müller, Europarecht, in: Kessler/Kröner/Köhler, Konzernsteuerrecht, 3. Aufl. 2018, 78, Rn. 225; Schönfeld, Praxisfragen, IStR 2012, 370; von Brocke, FG-Entscheidungen, DStR 2010, 966 f.; Maack/Kersten, GAV, DStR 2019, 2285 u. 2288.

[192] Vgl. Maack/Kersten, GAV, DStR 2019, 2289; a.A. Gosch, Border-Organschaften, IWB 2012, 697.

[193] Vgl. Maack/Kersten, GAV, DStR 2019, 2286-2288.

[194] Vgl. Homburg, Konzernverluste, IStR 2010, 251; von Brocke, FG-Entscheidungen, DStR 2010, 967; Homburg, I R 16/10 (Anm.), IStR 2011, 111; Glahe, Organschaft, IStR 2012, 133; von Brocke/Müller, Europarecht, in: Kessler/Kröner/Köhler, Konzernsteuerrecht, 3. Aufl. 2018, 78, Rn. 225; Maack/Kersten, GAV, DStR 2019, 2285; nach h.M. ist eine Verlustübernahmeverpflichtung jedoch möglich, vgl. etwa Kolbe in H/H/R, KStG, § 14 KStG, Anm. 12 (Januar 2015); Balmes/Brück/Ribbrock, Verlustnutzung, BB 2006, 188 f.; Englisch, C-446/03 (Anm.), IStR 2006, 23; Mitschke, Nulltarif, DStR 2010, 1371; Prinz/Ludwig, Gewinnabführungsvertrag, DB 2020, 1028; zum reduzierten GAV vgl. Rödder/Liekenbrock, in: Rödder/Herlinghaus/Neumann, KStG, § 14 KStG, Rz. 89 (2015); Krumm, in: Blümich, KStG, § 14 KStG Rz. 39 (August 2018); Micker, Verlustverrechnung, IWB 2013, 313 f.; Schnitger, Betriebsausgaben, IStR 2015, 773; Kessler/Arnold, Organschaft, IStR 2016, 231; Stöber,Verlustverrechnung, DStZ 2016, 589.

[195] Vgl. Gosch, Border-Organschaften, IWB 2012, 696 f.

[196] Vgl. Krumm, in: Blümich, KStG, § 14 KStG, Rn. 39 (August 2018).

[197] Vgl. Scheunemann, Verlustberücksichtigung, IStR 2006, 147 f.

[198] Vgl. Schleswig-Holsteinisches FG v. 13.03.2019, 1 K 218/15, Rev. eingelegt (Az. des BFH: I R 26/19), EFG 2019, 1466, Rn. 93.

[199] Vgl. Maack/Kersten, GAV, DStR 2019, 2288.

[200] Vgl. EuGH v. 19.06.2019, C-608/17, Rs. „Holmen“, EU:C:2019:511, Rn. 40; EuGH v. 19.06.2019, C-607/17, Rs. „Merima Holding“, EU:C:2019:510, Rn. 28.

[201] Vgl. Freimuth/Löprick, Verluste, StuB 2019, 737.

[202] Vgl. EuGH v. 31.02.2005, C-446/03, Rs. „Marks & Spencer“, EU:C:2005:763, Rn. 21; Herzig/Wagner, Gruppenbesteuerungssysteme, DStR 2006, 8; van Lishaut, Reform, FR 2009, 1036.

[203] Vgl. Scheunemann, Verlustberücksichtigung, IStR 2006, 148.

[204] Vgl. Herzig/Wagner, Gruppenbesteuerungssysteme, DStR 2006, 8; Scheunemann, Verlustberücksichtigung, IStR 2006, 149; van Lishaut, Reform, FR 2009, 1036.

[205] Vgl. BFH v. 29.10.1985, IX R 107/82, BStBl. II 1986, 217, Rz. 41; BFH v. 27.07.1999, VIII R 36/98, BStBl. II 1999, 769, Rz. 19; Scheunemann, Verlustberücksichtigung, IStR 2006, 150; Kleinert/Nagler, VI R 92/04 (Anm.), DB 2005, 2791.

[206] Vgl. Scheunemann, Verlustberücksichtigung, IStR 2006, 149 f.

Ist im Ansässigkeitsstaat der Tochtergesellschaft grundsätzlich kein Verlustabzug vorgesehen, soll wohl schon im Entstehungsjahr der Verluste die Finalität eintreten.[207] Gibt es im Staat der Tochtergesellschaft eine Frist, bis zu der die Verluste genutzt werden müssen und ab der der Verlustvortrag anderenfalls verfällt, soll dies nach Meinung des BFH nicht für eine Finalität der Verluste ausreichen.[208] Mussten die Verluste aufgrund einer beschränkten Verrechenbarkeit mit Gewinnen der Gesellschaft, bei der sie entstanden sind, vorgetragen werden oder durften sie nicht mit Gewinnen einer anderen Gesellschaft verrechnet werden, können dennoch hieraus finale Verluste entstehen, wenn die finale Versagung der Abzugsmöglichkeit vorliegt.[209] Können die Verluste im Jahr der Abwicklung einer Gesellschaft nicht übertragen werden, ist dies nicht relevant für die Entstehung finaler Verluste, die dennoch nur dann vorliegt, wenn die Muttergesellschaft nachweist, dass die Verluste auch in künftigen Zeiträumen nicht bei einem Dritten berücksichtigt werden können.[210] Ein freiwilliger Verzicht auf den Verlustausgleich im Ansässigkeitsstaat der Tochtergesellschaft soll nicht zur Abzugsmöglichkeit im Staat der Muttergesellschaft führen.[211] Auch führt ein Gesellschafterwechsel grds. nicht dazu, dass die Verluste im Inland abziehbar werden, da die Verluste i.d.R. vom Käufer genutzt werden können; eine Ausnahme hiervon wäre eine § 8c KStG entsprechende Mantelkaufregelung im Ansässigkeitsstaat der Tochtergesellschaft.[212] Bezogen auf Betriebsstätten hat der EuGH entschieden, dass bei einer Freistellung der Gewinne keine Pflicht zum Abzug finaler Betriebsstättenverluste besteht, unklar ist jedoch, ob dieses Urteil auch auf Tochtergesellschaften anzuwenden ist.[213] Abziehbar sind finale Verluste frühestens in dem Jahr, in dem sie final werden, ein rückwirkender Abzug findet also nicht statt.[214] Die Tatsache, dass durch diese zeitliche Verschiebung bei der Verlustberücksichtigung ein Liquiditätsnachteil für den Konzern entsteht, wurde vom EuGH nicht berücksichtigt.[215]

Wenn im Ansässigkeitsstaat der Tochtergesellschaft finale Verluste entstanden sind, ist schließlich noch zu klären, nach welchen Vorschriften die Höhe ermittelt wird. Eine Ermittlung nach dem ausländischen Handelsrecht und eine anschließende Anpassung nach den Normen des EStG bzw. KStG würde der Gesetzessystematik widersprechen, da dann auch das Maßgeblichkeitsprinzip der Handelsbilanz für die Steuerbilanz nach § 5 Abs. 1 S. 1 EStG gelten würde.[216]

Daher sind die Verluste nach h.M. für den Abzug bei dem Mutterunternehmen im Inland neu nach den steuerlichen Vorschriften im Inland zu ermitteln, um in dieser Hinsicht eine Gleichbehandlung zum Inlandssachverhalt herzustellen.[217] Zudem fordert der EuGH in dem Urteil eine Vergleichbarkeit zwischen inländischen und ausländischen Tochtergesellschaften, was dafür spricht, dass beide dieselben Gewinnermittlungsvorschriften anwenden müssen.[218] Dies führt allerdings zu der Problematik, dass die Verluste ggf. aufwendig für viele Jahre nach deutschem Steuerrecht zu ermitteln sind, wenn sie über einen langen Zeitraum entstanden sind.[219]

4.1.2. Doppelt ansässige Organgesellschaft

Rechtsstand

Doppelt ansässige Gesellschaften, d.h. solche mit Ort der GL im Inland und Sitz in einem Mitgliedstaat der EU oder einem Vertragsstaat des EWR, können nach § 14 Abs. 1 S. 1 KStG seit einer Gesetzesänderung 2013 grds. Organgesellschaft werden.[220] Nach ausländischem Recht gegründete Kapitalgesellschaften, wie es i.d.R. bei Kapitalgesellschaften mit Sitz in einem anderen Staat der Fall ist, können nur nach § 17 KStG Organgesellschaft werden, da sie nicht zu den in § 14 Abs. 1 S. 1 KStG genannten Gesellschaften gehören.[221] Der inländische Ort der GL stellt dabei die unbeschränkte Körperschaftsteuerpflicht nach § 1 Abs. 1 Nr. 1 KStG sicher.[222]

Tatsächlich ist es aber i.d.R. weiterhin nicht möglich, eine gewerbe- und körperschaftsteuerliche Organschaft mit einer ausländischen Tochtergesellschaft zu gründen, da der mit einer ausländischen Gesellschaft abgeschlossene GAV nicht von der Finanzverwaltung anerkannt wird.[223] Hierbei gibt es

[207] Vgl. Scheunemann, Verlustberücksichtigung, IStR 2006, 148.

[208] Vgl. BFH v. 09.06.2010, I R 100/09, BStBl. II 2010, 1065, Rz. 12; BFH v. 09.06.2010, I R 107/09, BFH/NV 2010, 1744, Rn. 20; gl.A. van Lishaut, Reform, FR 2009, 1036; a.A. Herzig/Wagner, Gruppenbesteuerungssysteme, DStR 2006, 8; Scheunemann, Verlustberücksichtigung, IStR 2006, 148.

[209] Vgl. EuGH v. 19.06.2019, C-608/17, Rs. „Holmen", EU:C:2019:511, Rn. 41-45.

[210] Vgl. EuGH v. 19.06.2019, C-608/17, Rs. „Holmen", EU:C:2019:511, Rn. 40.

[211] Vgl. Scheunemann, Verlustberücksichtigung, IStR 2006, 147 f.; Lang, Analyse, SWI 2006, 9; van Lishaut, Reform, FR 2009, 1036.

[212] Vgl. Scheunemann, Verlustberücksichtigung, IStR 2006, 149; Herzig/Wagner, Gruppenbesteuerungssysteme, DStR 2006, 8; van Lishaut, Reform, FR 2009, 1036.

[213] Vgl. EuGH v. 17.12.2015, C-388/14, Rs. „Timac Agro Deutschland", EU:C:2015:829, Rn. 65 f.; ablehnend Kessler/Arnold, Organschaft, IStR 2016, 230.

[214] Vgl. FG Rheinland-Pfalz v. 17.03.2010, 1 K 2406/07, rkr., EFG 2010, 1632, Rn. 93 f.; BFH v. 22.10.2010, I R 16/10, BFH/NV 2011, 524, Rn. 10; gl.A. van Lishaut, Reform, FR 2009, 1036.

[215] Vgl. Scheunemann, Schlussanträge, IStR 2005, 308; Herzig/Wagner, Gruppenbesteuerungssysteme, DStR 2006, 8; Wittkowski, Verlustverrechnung, 2008, 79; Herzig, Gruppenbesteuerung, StuW 2010, 216.

[216] Vgl. Scheunemann, Verlustberücksichtigung, IStR 2006, 150.

[217] Vgl. EuGH v. 13.12.2005, C-446/03, Rs. „Marks & Spencer", EU:C:2005:763, Rn. 22; Herzig/Wagner, Zukunft, DB 2004, 4; Herzig/Englisch/Wagner, Berücksichtigung, Der Konzern 2005, 312; Balmes/Brück/Ribbrock, Organschaftsbesteuerung, BB 2005, 968.; Scheunemann, Verlustberücksichtigung, IStR 2006, 150; Hey, Organschaft, GmbHR 2006, 115; van Lishaut, Reform, FR 2009, 1036.

[218] Vgl. EuGH v. 31.02.2005, C-446/03, Rs. „Marks & Spencer", EU:C:2005:763, Rn. 22; Scheunemann, Verlustberücksichtigung, IStR 2006, 150.

[219] Vgl. Frotscher, Organschaft, IStR 2011, 703.

[220] Vgl. § 14 Abs. 1 S. 1 KStG; Gesetz zur änderung und Vereinfachung der Unternehmensbesteuerung und des steuerlichen Reisekostenrechts v. 20.02.2013, BGBl. I 2013, 285.

[221] Vgl. §§ 14 Abs. 1 S. 1, 17 KStG; Dölker, in: Dölker/Ergenziger/Feldgen u.a., KStG, § 14 KStG, Rz. 29 (06.02.2020); Dötsch/Pung, Organschaft, DB 2013, 305.

[222] Vgl. OFD Frankfurt/M. v. 12.11.2019, S 2770 A - 55- St 55, DB 2020, 257.

[223] Vgl. Boller/Hackemann, Prüfstand, IStR 2020, 42; Prinz/Ludwig, Gewinnabführungsvertrag, DB 2020, 1022.

zwei Problemfelder: Durch das Erfordernis eines GAV nach § 291 Abs. 1 AktG wird auf einen deutschen Vertragstypus abgestellt, welcher nicht mit allen Rechtsordnungen vereinbar ist.[224] Zudem werden die Eintragungserfordernisse der Finanzverwaltung als zu hoch eingestuft, da das ausländische Register nicht als gleichwertig mit dem inländischen Register angesehen wird, sich die Tochtergesellschaft aufgrund ihres ausländischen Gesellschaftsstatuts jedoch nicht in das inländische HR eintragen lassen kann.[225] Aus diesem Grund hat die EU-Kommission am 25.07.2019 mit einem Aufforderungsschreiben an die Bundesregierung ein Vertragsverletzungsverfahren gegen die Bundesrepublik Deutschland eingeleitet,[226] da auch eine mittelbare Diskriminierung, wie es das Eintragungserfordernis sein könnte, dem Diskriminierungsverbot aufgrund der Ungleichbehandlung in der Praxis widerspräche.[227] Hintergrund dieses Vertragsverletzungsverfahrens war der Fall einer nach niederländischem Recht gegründeten Tochtergesellschaft mit Ort der GL in Deutschland, welche mit der deutschen Konzernmutter eine Organschaft begründen wollte. Zu diesem Zweck wurde ein GAV abgeschlossen, beim zuständigen niederländischen Registergericht hinterlegt sowie die Hinterlegung im niederländischen HR und infolge dessen auch im deutschen Register der Zweigniederlassung eingetragen.[228] Die deutsche Finanzverwaltung hielt jedoch die Form der Eintragung des GAVs in das niederländische HR für nicht ausreichend und forderte einen nach deutschem Recht abgeschlossenen GAV, wie er in § 291 Abs. 1 AktG beschrieben wird.[229]

Erneut ist hier die Frage zu klären, welches Recht für den GAV, in diesem Fall zwischen einer doppelt ansässigen Organgesellschaft und einem inländischen Mutterunternehmen, anzuwenden ist. Nach h.M. gilt für den GAV grds. nicht das Vertragsstatut, sondern das Gesellschaftsstatut, da er in seiner Wirkung einer Satzungsänderung entspricht und somit als Organisationsvertrag einzustufen ist.[230] Dieses richtet sich nach h.M., wie oben bereits ausgeführt, nach dem Gesellschaftsstatut der Untergesellschaft, da es auf ihrer Seite dem Rechtsverhältnis das typische Gepräge gibt.[231]

Bei einer doppelt ansässigen Organgesellschaft stellt sich allerdings die Frage, welches der beiden möglichen Rechtssysteme als ihr Gesellschaftsrecht anzusehen ist: Das Recht des Staates, in dem der Ort der GL liegt oder das Recht des Staates, in dem der Sitz liegt. Dies ist danach zu entscheiden, ob das jeweilige internationale Privatrecht die Sitz- oder die Gründungstheorie vertritt.[232] Deutschland verfolgt grds. die Sitztheorie, wonach das Recht des Staates anzuwenden ist, in dem die Gesellschaft ihren Verwaltungssitz, also ihren Ort der GL, hat.[233] Die Sitztheorie wurde jedoch vonseiten des EuGH mit Verweis auf die Niederlassungsfreiheit abgelehnt, sodass nun für Länder mit statuarischem Sitz in EU-/EWR-Staaten die Gründungstheorie anzuwenden ist.[234] Damit gilt für diese Gesellschaften das Recht des Staates, in dem sie gegründet wurden und registriert sind. Das Gesellschaftsstatut wird also nach internationalem Gesellschaftsrecht nach dem Sitzland der Gesellschaft bestimmt, was im Fall einer doppelt ansässigen Zuzugsgesellschaft das Land ihres Verwaltungssitzes ist.[235]

Gewinnabführungsvertrag

Weder das Gesetz noch die KStR geben Hinweise dazu, wie ein grenzüberschreitender GAV für eine Organschaft, also für die Verrechnung von mehr als nur finalen Verlusten, zu gestalten ist.[236] Die OFD Frankfurt/M. und das FinM SH haben sich nach Einleiten des Vertragsverletzungsverfahrens durch die EU-Kommission bemüht, diese Lücke durch je ein Rundschreiben zu schließen, in welchen sie die gleichen vier Bedingungen formulierten:[237]

Der GAV muss vollständig den Vorgaben des § 291 AktG entsprechen und insbesondere die Pflicht zur Verlustübernahme nach § 302 AktG beinhalten.[238] Die Pflicht zur Verlustübernahme nach § 302 AktG sollte dabei durch einen dynamischen Verweis umgesetzt werden, d.h. nicht durch ein Zitat des § 302 AktG, sondern durch einen Verweis auf § 302

[224]Vgl. Grotherr, Tatbestandsmerkmal, FR 1995, 11 f.; Grotherr, Unzulänglichkeiten, StuW 1995, 149; Herzig/Wagner, Zukunft, DB 2005, 5; Scheunemann, Verlustberücksichtigung, IStR 2006, 146; C. Witt, Konzernbesteuerung, FR 2009, 1048; Frotscher, Organschaft, IStR 2011, 702; Prinz/Ludwig, Gewinnabführungsvertrag, DB 2020, 1023; ausführlich dazu vgl. Scheunemann, Konzernbesteuerung, 2005, 121-128.

[225]Vgl. Prinz/Ludwig, Gewinnabführungsvertrag, DB 2020, 1024.

[226]Vgl. Europäische Kommission, Vertragsverletzungsverfahren v. 25.7.2019, 2019/4053, INF/19/4251, Impossibility to fulfil formal requirements for qualifying as a group company for German tax consolidation purposes.

[227]Vgl. EuGH v. 12.02.1974, C-152/73, Rs. „Sotgiu“, EU:C:1974:13, Rn. 11.

[228]Vgl. Boller/Hackemann, Prüfstand, IStR 2020, 42 f.

[229]Vgl. Boller/Hackemann, Prüfstand, IStR 2020, 43.

[230]Vgl. Koch, in: Hüffer/Koch, AktG, § 291 AktG, Rn. 17 (2020); Spahlinger/Wegen, Gesellschaftsrecht, 2005, 102; Kort, Beherrschungsvertrag, NZG 2009, 366; Scheunemann, Verlustberücksichtigung, IStR 2006, 146; Krüger/Epe, Fragen, in: Prinz/Witt, Organschaft, 2. Aufl. 2019, Rz. 26.3; Prinz/Ludwig, Gewinnabführungsvertrag, DB 2020, 1025; Müller/Detmering/Lieber, Organschaft, 11. Aufl. 2020, Rz. 193.

[231]Vgl. BFH v. 07.12.2011, I R 30/08, BStBl. II 2012, 507, Rn. 21; Bärwaldt/Schabacker, Unternehmensverträge, AG 1998, 186; Maul, Vollstreckungsfragen, AG 1998, 407 f.; Picot/Land, Unternehmenskauf, DB 1998, 1606; Scheunemann, Verlustberücksichtigung, IStR 2006, 146; van Lishaut, Reform, FR 2009, 1032; Frotscher, Organschaft, IStR 2011, 702; Glahe, Organschaft, IStR 2012, 131; Hoene, Gewinnabführungsvertrag, IStR 2012, 463; Kessler/Arnold, Organschaft, IStR 2016, 230; Koehler, Gewinnabführungsvertrag, Der Konzern 2018, 326; Krüger/Epe, Fragen, in: Prinz/Witt, Organschaft, 2. Aufl. 2019, 928 f., Rz. 26.4; a.A., wonach die Gesellschaftsstatute beider Gesellschaften gelten sollen: Einsele, Unternehmen, ZGR 1996, 50.

[232]Vgl. Koehler, Gewinnabführungsvertrag, Der Konzern 2018, 326.

[233]Vgl. EuGH v. 05.11.2002, C-208/00, Rs. „überseering BV“, EU:C:2002:632, Rn. 4; Schmidt-Leithoff, in: Rowedder/Schmidt-Leithoff, GmbHG, Einl., Rn. 269 u. 273 (2017); Koehler, Gewinnabführungsvertrag, Der Konzern 2018, 326.

[234]Vgl. EuGH v. 05.11.2002, C-208/00, Rs. „überseering BV“, EU:C:2002:632, Rn. 94 f.; Emmerich, in: Emmerich/Habersack, Konzernrecht, § 291 AktG, Rn. 33 (2019).

[235]Vgl. Brink, in: Schnitger/Fehrenbacher, KStG, § 14 KStG, Rn. 242 (2018); Schnitger, Fragestellungen, IStR 2013, 84; Krüger/Epe, Fragen, in: Prinz/Witt, Organschaft, 2. Aufl. 2019, Rz. 26.6.

[236]Vgl. Prinz/Ludwig, Gewinnabführungsvertrag, DB 2020, 1024.

[237]Vgl. OFD Frankfurt/M. v. 12.11.2019, S 2770 A - 55- St 55, DB 2020, 257; FinM SH v. 17.01.2020, VI 313-S 2770-077, DB 2020, 698.

[238]Vgl. OFD Frankfurt/M. v. 12.11.2019, S 2770 A - 55- St 55, DB 2020, 257; FinM SH v. 17.01.2020, VI 313-S 2770-077, DB 2020, 698.

AktG in seiner aktuellen Fassung.[239] Weiterhin muss der GAV nach dem ausländischen Recht zulässig sein.[240] Dies bedeutet nicht, dass das ausländische Recht den Vertragstypus eines GAV kennen muss, sondern lediglich, dass es dem Abschluss eines GAV nicht entgegenstehen darf, was nicht vom ausländischen Staat erzwungen werden kann.[241] Der GAV muss in eintragungspflichtiger Form vereinbart werden, d.h. es muss entweder eine Eintragungspflicht für den Vertrag in dem ausländischen Unternehmensregister bestehen oder die Regelungen für die Gewinnabführung werden in die Satzung der beherrschten Gesellschaft aufgenommen, wobei die Satzungsänderung eintragungspflichtig sein muss.[242] Schließlich muss der GAV, wenn er nicht in die Satzung selber aufgenommen wird, einen satzungsändernden und nicht nur einen satzungsüberlagernden Charakter haben.[243] Dies ist etwa bei einem deutschen GAV der Fall, da er in die Autonomie und in die Organisations- und Finanzverfassung der untergeordneten Gesellschaft eingreift.[244] Sind diese vier Bedingungen erfüllt, so leitet die Finanzverwaltung aus § 14 Abs. 1 S. 2 KStG ab, dass die Organschaft frühestens ab dem Jahr der Eintragung des GAV bzw. der Satzungsänderung in das ausländische Register anzuerkennen ist.[245]

Damit werden zwar nach ausländischem Recht abgeschlossene Gewinnabführungsverträge anerkannt,[246] bloße schuldrechtliche Verträge werden aber trotz der Verfügungen weiterhin nicht anerkannt, was vom Schrifttum kritisiert wird.[247] Ein weiterer Kritikpunkt ist die Tatsache, dass die aktuell geforderte Ausgestaltung in den meisten Mitgliedstaaten nicht erfüllbar ist.[248] Die meisten ausländischen Rechtsordnungen kennen das Rechtsinstitut des GAV nicht.[249] Dies ist allerdings grundsätzlich unproblematisch, da der ausländische GAV mit hoher Wahrscheinlichkeit ohnehin nicht dem inländischen GAV nach § 291 Abs. 1 AktG und somit dem in § 14 Abs. 1 S. 1 KStG geforderten Vertrag gleichen würde.[250] Relevant ist nur, dass das nationale Gesellschaftsrecht des Sitzlandes der Tochtergesellschaft dem Vertrag nicht entgegensteht.[251] Es gibt Länder im Gemeinschaftsgebiet, in denen eine solche Einflussnahme wie durch den GAV festgelegt nicht zulässig ist, etwa Irland.[252] In diesen Fällen ist der GAV unionsrechtlich als eine mittelbare Diskriminierung anzusehen.[253] Problematisch ist, dass, selbst wenn der Abschluss im Ausland möglich ist, einzelne Vorschriften des ausländischen Gesellschaftsrechts der Abführung des vollen Gewinns entgegenstehen können, etwa Begrenzungen zur Höhe von Ausschüttungen.[254] Denkbar ist auch, dass die Abführung des nach ausländischem Handelsrecht ermittelten Gewinns nicht dem ganzen Gewinn nach deutschem Handelsrecht entspricht und diese Voraussetzung daher als nicht erfüllt angesehen wird.[255] Hierbei könnte § 17 S. 2 Nr. 1 KStG eine Lösung bieten, wonach der abgeführte Gewinn lediglich den in § 301 AktG genannten Betrag nicht überschreiten darf.[256]

Aus der Literatur wird daher gefordert, dass, soweit es die ausländische Rechtsordnung zulässt, schuld- und organisationsrechtliche Elemente berücksichtigt werden sollen und der so abgeschlossene Vertrag anerkannt werden sollte.[257] Diese Anerkennung wäre womöglich durch bloße Auslegung auch ohne Gesetzesänderung zulässig.[258]

Das Eintragungserfordernis wird als diskriminierungsverdächtig kritisiert, da die meisten Länder keinen GAV kennen und aus diesem Grund i.d.R. keine Eintragungspflicht für diesen Vertragstypus fordern werden.[259] Aufgrund dieses möglichen Mangels an der Eintragungspflicht, die der Steuerpflichtige nicht zu verschulden hat, wird gefordert, das Erfordernis der Eintragungspflicht aufzugeben und nur an die tatsächliche Eintragung anzuknüpfen.[260] Beim GAV gehe es nicht um Formalitäten, sondern um das tatsächliche Tragen von Chancen und Risiken durch das Mutterunternehmen.[261] Das Eintragungserfordernis könne nur dann bestehen bleiben, wenn

[239] Vgl. Dötsch/Pung, Organschaft, DB 2013, 314.
[240] Vgl. OFD Frankfurt/M. v. 12.11.2019, S 2770 A - 55- St 55, DB 2020, 257; FinM SH v. 17.01.2020, VI 313-S 2770-077, DB 2020, 698.
[241] Vgl. Herzig/Wagner, Gruppenbesteuerungssysteme, DStR 2006, 9; Scheunemann, Verlustberücksichtigung, IStR 2006, 146.
[242] Vgl. OFD Frankfurt/M. v. 12.11.2019, S 2770 A - 55- St 55, DB 2020, 257; FinM SH v. 17.01.2020, VI 313-S 2770-077, DB 2020, 698.
[243] Vgl. OFD Frankfurt/M. v. 12.11.2019, S 2770 A - 55- St 55, DB 2020, 257; FinM SH v. 17.01.2020, VI 313-S 2770-077, DB 2020, 698.
[244] Vgl. Emmerich, in: Emmerich/Habersack, Konzernrecht, § 291 AktG, Rn. 52 (2019); Paschos, in: Henssler/Strohn, Gesellschaftsrecht, § 291 AktG, Rn. 4 (2019); Kort, Beherrschungsvertrag, NZG 2009, 366; Link, Grundlagen, in: Prinz/Witt, Organschaft, 2. Aufl. 2019, 58, Rz. 2.20; Prinz/Ludwig, Gewinnabführungsvertrag, DB 2020, 1025.
[245] Vgl. OFD Frankfurt/M. v. 12.11.2019, S 2770 A - 55- St 55, DB 2020, 257; FinM SH v. 17.01.2020, VI 313-S 2770-077, DB 2020, 698.
[246] Vgl. Dötsch/Pung, in: D/P/M, KSt, § 14 KStG, Tz. 98 (Juni 2020).
[247] Vgl. Schober, in: Musil/Weber-Grellet, Steuerrecht, § 14 KStG, Rn. 206 (2019); Weinberger, Gewinnabführungsvertrag, Der Konzern 2020, 63; Prinz/Ludwig, Gewinnabführungsvertrag, DB 2020, 1024.
[248] Vgl. Dötsch/Pung, Organschaft, DB 2013, 306; Weinberger, Gewinnabführungsvertrag, Der Konzern 2020, 63.
[249] Vgl. Olbing, in: Streck, KStG, § 14 KStG, Rz. 12 (2018); Dötsch/Pung, in: D/P/M, KSt, § 14 KStG, Tz. 339a (Februar 2020); Dölker, in: Dölker/Ergenziger/Feldgen u.a., KStG, § 14 KStG, Rz. 29 (06.02.2020); Hommelhoff, in: Lutter/Hommelhoff, GmbHG, Anhang zu § 13, Rz. 96 (2020); Scheunemann, Verlustberücksichtigung, IStR 2006, 146; Raupach, Quelle, in: Hommelhoff/Rawert/Schmidt, FS Priester, 2007, 638; van Lishaut, Reform, FR 2009, 1032.
[250] Vgl. Prinz/Ludwig, Gewinnabführungsvertrag, DB 2020, 1023.
[251] Vgl. Liebscher, in: Fleischer/Goette, GmbHG, Anhang zu § 13 GmbHG, Rn. 682 (2018); Frotscher, Organschaft, IStR 2011, 702; Boller/Hackemann, Prüfstand, IStR 2020, 44; a.A. Dötsch, in: D/P/M, KSt, § 14 KStG, Tz. 13 (August 2020) u. Glahe, Organschaft, IStR 2012, 131, wonach der andere Staat den GAV kennen muss; Balmes/Brück/Ribbrock, Organschaftsbesteuerung, BB 2005, 969, wonach ein grenzüberschreitender Gewinnabführungsvertrag nach deutschem Gesellschaftsrecht nicht abgeschlossen werden könne.
[252] Vgl. Prinz/Ludwig, Gewinnabführungsvertrag, DB 2020, 1026.
[253] Vgl. Hey, Organschaft, GmbHR 2006, 118; van Lishaut, Reform, FR 2009, 1032; kritisch Herzig, Gruppenbesteuerung, StuW 2010, 216.
[254] Vgl. Herzig/Wagner, Gruppenbesteuerungssysteme, DStR 2006, 9.
[255] Vgl. Dötsch/Pung, in: D/P/M, KSt, § 14 KStG, Tz. 99 (Juni 2020).
[256] Vgl. Dötsch/Pung, in: D/P/M, KSt, § 14 KStG, Tz. 99 (Juni 2020).
[257] Vgl. Kessler/Arnold, Organschaft, IStR 2016, 231; Weinberger, Gewinnabführungsvertrag, Der Konzern 2020, 63; Prinz/Ludwig, Gewinnabführungsvertrag, DB 2020, 1026.
[258] Vgl. Weinberger, Gewinnabführungsvertrag, Der Konzern 2020, 63.
[259] Vgl. Weinberger, Gewinnabführungsvertrag, Der Konzern 2020, 63; Prinz/Ludwig, Gewinnabführungsvertrag, DB 2020, 1024.
[260] Vgl. Prinz/Ludwig, Gewinnabführungsvertrag, DB 2020, 1029; Weinberger, Gewinnabführungsvertrag, Der Konzern 2020, 63.
[261] Vgl. Prinz/Ludwig, Gewinnabführungsvertrag, DB 2020, 1024.

es für die Wirksamkeit des Vertrages erforderlich ist.[262]

Insgesamt wird das Erfordernis des GAV nach h.M. als klar unionsrechtswidrig eingestuft.[263] Aus diesem Grund wird einem vom Steuerpflichtigen veranlassten Rechtsbehelfsverfahren in solch einem Zweifelsfall gute Chancen eingeräumt.[264]

Die Bundesregierung hat deutlich gemacht, dass sie von ihrem Standpunkt nicht abrücken wird, da sie in der aktuellen Regelung keinen Verstoß gegen Unionsrecht erkennen könne.[265] Folglich ist vom Gesetzgeber nicht geplant, das AktG oder die Ertragsteuergesetze zu ändern, um den GAV oder sein Eintragungserfordernis abzuschaffen oder formelle Anforderungen abzuschwächen.[266]

Wegzugsgesellschaft und Gesellschaft mit inländischer Betriebsstätte

Eine Wegzugsgesellschaft, d.h. eine Gesellschaft, die ihren Sitz im Inland hat, ihren Ort der GL jedoch in das Ausland verlegt hat, kann nicht Organgesellschaft sein, da sie das Erfordernis des § 14 Abs. 1 S. 1 KStG nicht erfüllt, obwohl sie aufgrund des inländischen Sitzes unbeschränkt steuerpflichtig ist.[267] Der Abschluss des GAV wäre in diesem Fall zwar unproblematisch, da die Gesellschaft dann dem deutschen Gesellschaftsstatut unterläge, der fehlende Ort der GL im Inland macht jedoch die Organschaft in diesem Fall unmöglich.[268] Dieser Fall wäre etwa bei einer GmbH möglich, welche ihren Verwaltungssitz in einen anderen EU- oder EWR-Staat verlegt hat.[269] Die Tatsache, dass bei einer Wegzugsgesellschaft mit einem Mutterunternehmen mit inländischer Betriebsstätte keine Organschaft möglich ist, wird teilweise als Verstoß gegen die Niederlassungsfreiheit[270] und teilweise als unionsrechtlich zulässige Inländerdiskriminierung gesehen.[271] Eine Klage vor dem EuGH hätte also offene Erfolgschancen.

Ebenfalls nicht als Organgesellschaft qualifiziert ist eine Kapitalgesellschaft mit ausländischem Sitz und ausländischem Ort der GL, aber inländischer Betriebsstätte, etwa einer eingetragenen Zweigniederlassung.[272] Auch dies ist unionsrechtlich kritisch zu betrachten, da eine Ungleichbehandlung zwischen einer inländischen Kapitalgesellschaft mit inländischer Betriebsstätte und einer ausländischen Kapitalgesellschaft mit inländischer Betriebsstätte vorliegen könnte.[273] Denkbar wäre auch ein Verstoß gegen das Diskriminierungsverbot in Art. 24 Abs. 5 OECD-MA.[274]

4.2. Drittland

Die Organschaft ist aufgrund der Ansässigkeitserfordernisse für die Organgesellschaft (Sitz in einem EU- oder EWR-Staat, Ort der GL in Deutschland) nicht für eine Organgesellschaft mit Drittlandsbezug, d.h. für eine Gesellschaft mit Sitz und bzw. oder Ort der GL im Drittland, geöffnet.[275] Grundsätzlich können sich Gesellschaften aus Drittländern auch nicht auf die Niederlassungsfreiheit berufen, da die unionsrechtlichen Verträge räumlich nur beschränkt auf das Gemeinschaftsgebiet anzuwenden sind.[276] Zwar gibt es vereinzelt die Meinung, dass Entscheidungen des EuGH etwa bezogen auf finale Verluste auch einen Verstoß gegen die Kapitalverkehrsfreiheit beinhalten und daher gem. Art. 63 AEUV auch auf Drittstaaten anzuwenden sind.[277] Allerdings findet selbst in dem Fall, dass diese Sichtweise zutrifft, Art. 63 AEUV keine Anwendung auf die deutsche Organschaft, da dieser Artikel gem. Art. 64 Abs. 1 AEUV nicht auf Beschränkungen anzuwenden ist, die am 31.12.1993 aufgrund einzelstaatlicher oder gemeinschaftlicher Rechtsvorschriften für den Kapitalverkehr mit Drittländern im Zusammenhang mit Direktinvestitionen bestehen.[278] Dieser Bestandsschutz gilt auch für die Organschaft, da sie trotz diverser änderungen seit dem Jahr 1994 in Kern und Kernvoraussetzungen identisch geblieben ist.[279]

[262] Vgl. Kessler/Arnold, Organschaft, IStR 2016, 232.

[263] Vgl. Herzig/Wagner, Zukunft, DB 2005, 5; Balmes/Brück/Ribbrock, Organschaftsbesteuerung, BB 2005, 969; Raupach/Pohl, Konzernsteuerrecht, NZG 2005, 492; Hey, Organschaft, GmbHR 2006, 118; Linn/Reichl/Wittkowski, Verlustverrechnung, BB 2006, 631; Pache/Englert, Verlustverrechnung, IStR 2007, 49 f.; S. Witt, Organschaft, Ubg 2010, 741 f.; Mitschke, Ergebnisabführungsvertrag, IStR 2011, 185; Kahle/Schulz/Vogel, Reformvorschläge, Ubg 2011, 765; Kahle/Cortez, Zuzug, FR 2014, 685; Kessler/ Arnold, Organschaft, IStR 2016,231; Prinz/Ludwig, Gewinnabführungsvertrag, DB 2020, 1027.

[264] Vgl. Prinz/Ludwig, Gewinnabführungsvertrag, DB 2020, 1024.

[265] Vgl. BT-Drucks. 19/18624, 2.

[266] Vgl. BT-Drucks. 19/18624, 3 f.

[267] Vgl. § 14 Abs. 1 Nr. 1 KStG; Meilicke, Neuregelung, DB 2002, 911; Kessler/Arnold, Organschaft, IStR 2016, 229; Prinz/Ludwig, Gewinnabführungsvertrag, DB 2020, 1022.

[268] Vgl. Schnitger, Fragestellungen, IStR 2013, 86.

[269] Vgl. § 4a GmbHG; Prinz/Ludwig, Gewinnabführungsvertrag, DB 2020, 1022.

[270] Vgl. Schwenke, Organschaft, ISR 2013, 43; Schnitger, Fragestellungen, IStR 2013, 84 f.; Micker, Inlandsbezug, IWB 2013, 316 f.Jesse, Holding, in: Lutter/Bayer, Holding-Handbuch, 5. Aufl. 2015, 946 f., Rz. 14.541; Kessler/Arnold, Organschaft, IStR 2016, 229; Maack/Kersten, GAV, DStR 2019, 2284;

[271] Vgl. BFH v. 15.07.2005, I R 21/04, BStBl. II 2005, 716, Rz. 11; Dötsch/Pung, in: D/P/M, KSt, § 14 KStG, Tz. 96 (Juni 2020); Dötsch/Pung, Organschaft, DB 2013, 306.

[272] Vgl. Keß, in: L/S, GewStG, § 2 GewStG, Rz. 3550 (Oktober 2017); Dötsch/Pung, in: D/P/M, KSt, § 14 KStG, Tz. 96 (Juni 2020); Kessler/Arnold, Organschaft, IStR 2016, 229.

[273] Vgl. EuGH v. 06.09.2012, C-18/11, Rs. „Philips Electronics UK“, EU:C:2012:532, Leitsatz; Keß, in: L/S, GewStG, § 2 GewStG, Rz. 3550 (Oktober 2017); Dötsch/Pung, in: D/P/M, KSt, § 14 KStG, Tz. 96 (Juni 2020); Stöber, Verlustverrechnung, DStZ 2016, 590; Kessler/Arnold, Organschaft, IStR 2016, 229.

[274] Vgl. Schwenke, Organschaft, ISR 2013, 43.

[275] Vgl. Prinz/Ludwig, Gewinnabführungsvertrag, DB 2020, 122.

[276] Vgl. Scheunemann, Verlustberücksichtigung, IStR 2006, 152; Kessler/Arnold, Organschaft, IStR 2016, 229.

[277] Vgl. Scheunemann, Verlustberücksichtigung, IStR 2006, 152.

[278] Vgl. Art. 64 Abs. 1 AEUV.

[279] Vgl. §7a KStG i.d.F. v. 1969; Gesetz zur Senkung der Steuersätze und zur Reform der Unternehmensbesteuerung (Steuersenkungsgesetz - StSenkG) v. 23.10.2000, BGBl. I 2000, 1433; Gesetz zur Bekämpfung von Steuerverkürzungen bei der Umsatzsteuer und zur änderung anderer Steuergesetze (Steuerverkürzungsbekämpfungsgesetz - StVBG) v. 19.12.2001, BGBl. I 2001, 3922; UntStFG v. 20.12.2001, BGBl. I 2001, 3858; Gesetz zum Abbau von Steuervergünstigungen und Ausnahmeregelungen (Steuervergünstigungsabbaugesetz - StVergAbG) v. 16.5.2003, BGBl. I 2003, 660; Scheunemann, Verlustberücksichtigung, IStR 2006, 152 f.

In Einzelfällen könnte zwar ein Verstoß gegen ein DBA-rechtliches Diskriminierungsverbot vorliegen,[280] hierzu sind jedoch keine zustimmenden Gerichtsurteile im Zusammenhang mit dem aktuellen Gesetzeswortlaut bekannt.

5. Fazit

Insgesamt gibt es, je nach Sachverhalt, bei dem Versuch, eine grenzüberschreitende Organschaft zu errichten, verschiedene Hürden, die es aus Sicht des Steuerpflichtigen zu überwinden gilt. Bei einem Inbound-Fall, in dem das Mutterunternehmen eine inländische Betriebsstätte hat, ist vor allem darauf zu achten, dass die Beteiligung der Betriebsstätte zuzuordnen ist. Liegt keine inländische Betriebsstätte vor, ist keine Organschaft zwischen Schwestergesellschaften möglich, obwohl dies bei einem im EU-/EWR-Ausland ansässigen Mutterunternehmen nach h.M. klar unionsrechtswidrig ist.

Die Outbound-Fallkonstellation birgt zwar mit dem Richterrecht zu finalen Verlusten Hürden, welche aber für den Steuerpflichtigen recht klar definiert sind, wodurch es ihm möglich ist, entsprechende Gestaltungen vorzunehmen. Im Fall einer Organgesellschaft mit Ort der GL im Inland und Sitz in einem EU- oder EWR-Staat bleibt abzuwarten, welche weiteren Schritte die EU-Kommission aufgrund des Erfordernisses des GAV einleitet. Nach erfolgter übermittlung des Aufforderungsschreibens kann die EU-Kommission im nächsten Schritt eine mit Gründen versehene Stellungnahme abgeben, in welcher sie Deutschland dazu auffordert, eine übereinstimmung mit dem Unionsrecht herzustellen.[281] Kommt Deutschland dem nicht nach, kann der EuGH mit dem Sachverhalt beschäftigt werden, wobei dessen Urteil Folge zu leisten ist.[282] Ist die Tochtergesellschaft in einem Drittland ansässig, gibt es keine realistische Möglichkeit, eine Organschaft zu bilden.

Insgesamt sind die Regelungen zur grenzüberschreitenden Organschaft durchdrungen von sicher oder möglicherweise unionsrechtswidrigen Voraussetzungen. Der deutsche Gesetzgeber handelt dennoch eher zurückhaltend und scheint nicht proaktiv im Vorgriff auf mögliche Verfahren vor dem EuGH oder Vertragsverletzungsverfahren die Regelungen ändern zu wollen. Dies könnte zu einem unerwarteten Verlust von Besteuerungssubstrat führen, wenn der EuGH eine Entscheidung gegen eine der aktuellen Regelungen trifft.[283] Zudem würde eine umfassende Reform Rechtssicherheit für den Steuerpflichtigen bringen, ein Ende der häufigen änderungen des Gesetzestextes ermöglichen[284] und könnte durch großzügigere Ausgestaltung zu einem Standortvorteil für Deutschland führen.[285] Da eine Reform der Regelung aber nicht abzusehen ist, wird es wohl für unbestimmte Zeit bei dem - nicht zufriedenstellenden - Status quo bleiben.

[280] Vgl. Kessler/Arnold, Organschaft, IStR 2016, 229.
[281] Vgl. Barelt/Geberth, EU-Kommission, DB 2019, 1990.
[282] Vgl. Barelt/Geberth, EU-Kommission, DB 2019, 1990.
[283] Vgl. Kessler/Arnold, Organschaft, IStR 2016, 233.
[284] Vgl. Schnitger, Organschaftsreform, IStR 2014, 591.
[285] Vgl. Herzig, Gruppenbesteuerung, StuW 2010, 216; Kessler/Arnold, Organschaft, IStR 2016, 233.

Literatur

Altrichter-Herzberg, T. & Nuernberger, E. (2006). Verlustnutzung über die Grenze. *GmbHR*, 466–469.

Balmes, F., Brück, M. & Ribbrock, M. (2005). Der EuGH-Fall Marks & Spencer: Rückschlüsse für die deutsche Organschaftsbesteuerung. *BB*, 966–970.

Balmes, F., Brück, M. & Ribbrock, M. (2006). Das EuGH-Urteil Marks & Spencer: Grenzüberschreitende Verlustnutzung kommt voran! *BB*, 186–189.

Balmes, F., Grammel, R. & Sedemund, J. (2006). Berücksichtigung von Betriebsstättenverlusten trotz Freistellungsmethode. *BB*, 1474–1480.

Bartelt, M. & Geberth, G. (2019). EU-Kommission: Deutschland soll Gewinnabführungs- und Verlustübernahmeverträge anerkennen, die nach den Rechtsvorschriften eines anderen EU- oder EWR-Mitgliedstaatsgeschlossen wurden. *DB*, 1990.

Behrens, S. (2005, Dezember). EuGH rettet die deutsche Organschaft; Urteil in der Rechtssache Marks & Spencer verlangt aber eine Anpassung von deutschen Vorschriften. *Börsen-Zeitung* (244), 6.

Benecke, A. & Schnitger, A. (2013). Wichtige Änderungen bei der körperschaftsteuerlichen Organschaft durch das UntStG 2013. *IStR*, 143–157.

Blümich, W. (2020). Einkommensteuergesetz, Körperschaftsteuergesetz, Gewerbesteuergesetz, Kommentar, Loseblatt, München 2020 (Stand: 151. Erg.-Lfg. März 2020).

Blumenberg, J. & Hundeshagen, C. (2019). Gestaltungsfragen der ertragsteuerlichen Organschaft. In *Steuerliche Organschaft* (2. Aufl., S. 333–375). München: Prinz, U. and Witt, S.

Blumers, W. & Goerg, D. (2003). Gewerblichkeit von Organträger-Personengesellschaften. *BB*, 2203–2207.

Boller, T. & Hackemann, T. (2020). Die ertragsteuerliche Organschaft erneut auf dem Prüfstand der EU-Kommission. *IStR*, 41–46.

Bott, H. & Walter, W. (1996 (145. Erg.-Lfg. Mai 2020)). Körperschaftsteuergesetz, Kommentar. *Loseblatt*.

Breuninger, G. (2019). Grundlagen internationaler Organschaft. In *Steuerliche Organschaft* (2. Aufl., S. 909–926). Prinz, U. and Witt, S.

Bärwaldt, R. & Schabacker, J. (1998). Wirksamkeitserfordernisse grenzüberschreitender Unternehmensverträge i. S. d. § 291 AktG. *AG*, 182–189.

Cloer, A. & Kahlenberg, C. (2014). Die Gruppenbesteuerung in der EU im Lichte der EuGH-Rechtsprechung. *SteuK*, 511–514.

Cloer, A. & van Lück, N. (2019). Kein Abzug finaler Verluste bei Organschaft über die Grenze. *DStRK*, 295–295.

Dölker, A., Ergenzinger, P. & Feldgen, R. u. (2019, März). *KStG eKommentar*. Zugriff auf www.first.stollfuss.de/stotax

Dörr, I. (2004). Abschaffung oder Erweiterung der Organschaft?! zu den möglichen Konsequenzen der Rechtssache „Marks and Spencer plc". *IStR*, 265–272.

Dötsch, E. (2014). Organträger-Personengesellschaft mit ausländischen Gesellschaftern: Zur Anwendung des § 14 Abs. 1 Satz 1 Nr. 2 Satz 7 KStG. *DB*, 1215–1218.

Dötsch, E. & Pung, A. (2013). Gesetz zur Änderung und Vereinfachung der Unternehmensbesteuerung und des steuerlichen Reisekostenrechts: Die Änderungen bei der Organschaft. *DB*, 305–314.

Dötsch, E. & Pung, A. (2019). Zuordnung von Kapitalbeteiligungen zu einer inländischen Betriebsstätte einer PersGes. bei Vorhandensein ausländischer Mitunternehmer. *DB*, 146–148.

Dötsch, E., Pung, A. & Möhlenbrock, R. (2020). Die Körperschaftsteuer, Kommentar, Loseblatt, Stuttgart 2020 (Stand: 99. Erg.-Lfg. Juni 2020).

Einsele, D. (1996). Kollisionsrechtliche Behandlung des Rechts verbundener Unternehmen. *ZGR*, 40–54.

Emmerich, V. & Habersack, M. (2019). *Aktien- und GmbH-Konzernrecht, Kommentar* (9. Aufl.). München.

Endres, D. & Thies, A. (2002). Bekämpfung doppelter Verlustnutzung im UntStFG – Oder: Wer danebenschießt, muss eine Strafrunde laufen. *RIW*, 275–280.

Englisch, J. C. (2006). Anmerkung zum Urteil des EuGH v. 13.12.2005, C-446/03. *IStR*, 19–23.

Füger, R. (2003). Steuerplanerische Überlegungen nach den Änderungenbei der ertragsteuerlichen Organschaft. *BB*, 1755–1760.

Fleischer, H. & Goette, W. (2018). *Münchener Kommentar GmbHG, Kommentar* (3. Aufl.). München.

Freimuth, H. & Löprick, N. (2019). Berücksichtigung finaler Verluste im Kapitalgesellschaftskonzern. *StuB*, 734–737.

Frotscher, G. (2011). Grenzüberschreitende Organschaft – wo stehen wir? *IStR*, 697–703.

Frotscher, G. & Drüen, K. (2007 (Stand: 152.Erg.-Lfg. Januar 2020)). Kommentar zum Körperschaft- und Umwandlungsteuergesetz, Kommentar. *Loseblatt*.

Förster, G. (2003). Die Änderungen durch das StVergAbG bei der Einkommensteuer und der Körperschaftsteuer. *DB*, 899–905.

Glahe, M. (2012). Grenzüberschreitende Organschaft ohne Gewinnabführungsvertrag. *IStR*, 128–134.

Glanegger, P. & Güroff, G. (2017). *GewStG, Kommentar* (9. Aufl.). München.

Goebel, S. & Ungemach, M. (2013). Neuregelungen bei der Besteuerung ertragsteuerrechtlicher Organschaften mit Auslandsbezug. *NWB*, 595–605.

Goette, W. & Habersack, M. (2017). *Münchener Kommentar zum Aktiengesetz, Kommentar* (4. Aufl.). München.

Gosch, D. (2012). Über Cross Border-Organschaften - Abkehr des BFH vom strikten Inlandsbezug. *IWB*, 694–707.

Gosch, D. (2015). *Körperschaftsteuergesetz, Kommentar* (3. Aufl.). München.

Gosch, D. (2018). „Poolen binnen un buten" – Der RFH, der BFH und die Gewinngemeinschaft. *IStR*, 747–756.

Graw, C. (2010). Abzug finaler Verluste ausländischer Tochtergesellschaften bei der inländischen Muttergesellschaft. *DB*, 2469–2473.

Grotherr, S. (1993). Zur gegenwärtigen Bedeutung der Organschaft in der Konzernsteuerplanung. *BB*, 1986–2001.

Grotherr, S. (1995a). Der Abschluß eines Gewinnabführungsvertrags als (un-)verzichtbares Tatbestandsmerkmal der körperschaftsteuerlichen Organschaft. *FR*, 1–14.

Grotherr, S. (1995b). Unzulänglichkeiten des gegenwärtigen Organschaftskonzepts. *StuW*, 124–150.

Hackemann, T. (2013). Group Taxation in the European Union. In *Allocating Taxing Powers within the European Union* (S. 123–130). Berlin: Richelle, I. and Schön, W. and Traversa, E.

Heckerodt, D. (2019). Finale Betriebsstättenverluste: Kein Mythos, sondern Realität! *IStR*, 171–177.

Henkel, U. (2018). Beteiligung an ausländischen Kapitalgesellschaften. In *Steuerrecht international tätiger Unternehmen* (5. Aufl., S. 849–888). Köln: Mössner, J. and Baumhoff, H. and Greif, M.

Henssler, M. & Strohn, L. (2019). *Gesellschaftsrecht, Kommentar* (4. Aufl.). München.

Herbener, R. (2019). Holding als Organträger. In *Steuerliche Organschaft* (S. 673–695). München: Prinz, U. and Witt, S.

Herrmann, C., Heuer, G. & Raupach, A. (2018). EStG/KStG, Kommentar, Loseblatt, Köln 2018 (Stand: 296. Erg.-Lfg., Februar 2020).

Herzig, N. (2010). Die Zukunft der Gruppenbesteuerung. *StuW*, 214–231.

Herzig, N., Englisch, J. & Wagner, T. (2005). Steuerliche Berücksichtigung von Verlusten ausländischer Konzerntöchter. *U*, 298–318.

Herzig, N. & Wagner, T. (2005). Zukunft der Organschaft im EG-Binnenmarkt. *DB*, 1–9.

Herzig, N. & Wagner, T. (2006). EuGH-Urteil „Marks & Spencer" - Begrenzter Zwang zur Öffnung nationaler Gruppenbesteuerungssysteme für grenzüberschreitende Sachverhalte. *DStR*, 1-12.

Heurung, R. & Klübenspies, M. (2003). Ertragsteuerliche Organschaften im Lichte des BMF-Schreibens vom 26. 8. 2003. *BB*, 2483–2495.

Heurung, R., Schmidt, S. & Kraft, M. (2018). BB-Rechtsprechungsreport zur ertragsteuerlichen Organschaft 2017. *BB*, 470–476.

Hey, J. (2006). Die EuGH-Entscheidung in der Rechtssache Marks & Spencer und die Zukunft der deutschen Organschaft. *GmbHR*, 113–123.

Hüffer, U. & Koch, J. (2018). *Aktiengesetz, Kommentar* (13. Aufl.). München.

Hölters, W. (2017). *Aktiengesetz, Kommentar* (3. Aufl.). München.

Hoene, P. (2012). Der grenzüberschreitende Gewinnabführungsvertrag. *IStR*, 462–464.

Homburg, S. (2010). Die unheimliche Nummer Sechs – Eine Entscheidung zum Ausgleich grenzüberschreitender Konzernverluste. *IStR*, 246–252.

Homburg, S. (I R 16/10 (Anm.)). (2011). Anmerkung zum Urteil des BFH v. 9. 11. 2010, i R 16/10. *IStR*, 110–112.

Hommelhoff, P., Rawert, P. & Schmidt, K. (2007). *Festschrift für Hans-Joachim*

Priester zum 70. Geburtstag.

IDW. (2014). Zweifelsfragen zu den Neuerungen bei der ertragsteuerlichen Organschaft nach dem Gesetz zur Änderung und Vereinfachung der Unternehmensbesteuerung. *Ubg*, 213–216.

IDW. (2020, Mai). *IDW Stellungnahme zum Entwurf einer Allgemeinen Verwaltungsvorschrift über die Neufassung der Körperschaftsteuer-Richtlinien 2015 (KStR 2015)*. Zugriff auf https://www.idw.de/blob/83816/49b57b68f1bd41c9a34dc510db3f963c/down-bmf-kstr-e2015-anl2-data.pdf

Ismer, R. (2012). Gruppenbesteuerung statt Organschaft im Ertragsteuerrecht? *DStR*, 821–829.

Jesse, L. (2013). Neuregelungen zur ertragsteuerlichen Organschaft (Teil i). *FR*, 629–640.

Jesse, L. (Holding). (2015). § 14 Die nationale Holding aus steuerrechtlicher Sicht. In *Holding-Handbuch - Konzernrecht - Konzernsteuerrecht - Konzernarbeitsrecht – Betriebswirtschaft* (5. Aufl., S. 653–1003). Lutter, M. and Bayer, W.

Juretzek, P. (C-41/13 (Anm.)). (2014). Anmerkung zum Urteil des EuGH v. 12.06.2014, C-41/13. *GWR*, 336–336.

Kahle, H. & Cortez, B. (2014). Zuzug von Kapitalgesellschaften im Ertragsteuerrecht. *FR*, 673–687.

Kahle, H., Vogel, T. & Schulz, S. (2011). Internationale Aspekte der Organschaft unter besonderer Berücksichtigung aktueller Reformvorschläge. *Ubg*, 761–771.

Kempf, A., Loose, T. & Oskamp, M. (2018). Ist die Verlustbeschränkung des § 14 Abs. 1 S. 1 Nr. 5 KStG für die Gewerbesteuer relevant? *DStR*, 2511–2514.

Kessler, W. & Arnold, N. (2016). National begrenzte Organschaft. *IStR*, 226–233.

Kessler, W., Kröner, M. & Köhler, S. (2018). *Konzernsteuerrecht* (3. Aufl.). München.

Kleinert, J., Nagler, J. & Rehm, H. (2005). Gewinnbesteuerung nach 'Art des Hauses' mittels grenzüberschreitender Organschaft. *DB*, 1969–1875.

Kleinert, J. and Nagler, J. (VI R 92/04 (Anm.)). (2005). Anmerkung zum Urteil des BFH v. 23.12.2005, VI R 92/04. *DB*, 2788–2793.

Kley, M., Sünner, E. & Willemsen, A. (1997). *Festschrift für Wolfgang Ritter*.

Koehler, C. (2018). Der grenzüberschreitende Gewinnabführungsvertrag mit einer ausländischen GmbH. *Der Konzern*, 325–334.

Kokott, J. (2018). *Das Steuerrecht der Europäischen Union*. München.

Kort, M. (2009). Anwendung der Grundsätze der fehlerhaften Gesellschaft auf einen „verdeckten" Beherrschungsvertrag? *NZG*, 364–368.

Kraft, C. & Hohage, U. (2019). Abzugsverbot für Sonderbetriebsausgaben nach § 14 Abs. 1 Satz 1 Nr. 5 KStG oder § 4i EStG in Abhängigkeit vom ausländischen Gruppenbesteuerungsregime. *Ubg*, 305–315.

Krenzler, H., Herrmann, C. & Niestedt, M. (2019). EU-Außenwirtschafts- und Zollrecht, Kommentar, München 2019 (Stand: 14. Erg.-Lfg,. Oktober 2019) .

Krüger, A. & Epe, D. (2019). Gesellschaftsrechtliche Fragen. In *Steuerliche Organschaft* (2. Aufl., S. 927–941). Prinz, U. and Witt, S.

Krieger, G., Lutter, M. & Schmidt, K. (2013). *Festschrift für Michael Hoffmann-Becking zum 70. Geburtstag*.

Lang, M. (2006). Marks & Spencer – Eine erste Analyse des EuGH-Urteils. *SWI*, 3–12.

Lenski, E. & Steinberg, W. (1995 (130. Erg.-Lfg. Februar 2020)). Gewerbesteuergesetz, Kommentar. *Loseblatt*.

Lieber, B. (C-39/13, C-40/13, C-41/13 (Anm.)). (2014). Anmerkung zum Urteil des EuGH v. 12.06.2014, C-39/13, C-40/13, C-41/13. *ISR*, 278–280.

Link, S. (2019). Gesellschaftsrechtliche Grundlagen der Organschaft. In *Steuerliche Organschaft* (2. Aufl.). Köln: Prinz, U. and Witt, S.

Linn, A., Reichl, A. & Wittkowski, A. (2006). Grenzüberschreitende Verlustverrechnung: Möglichkeiten und Grenzen. *BB*, 630-638.

Lutter, M. & Bayer, W. (2015). *Holding-Handbuch - Konzernrecht - Konzernsteuerrecht - Konzernarbeitsrecht – Betriebswirtschaft* (5. Aufl.). Köln.

Lutter, M. & Hommelhoff, P. (2020). *GmbH-Gesetz Kommentar, Kommentar* (20. Aufl.). Köln.

Löwenstein, U. & Maier, J. (2002). Organschaft und eingeschränkte Verlustnutzung bei doppelt ansässigen Organträgern - Mitwirkung von Dr. Jekyll und Mr. Hyde bei den Neuregelungen des UntStFG? *IStR*, 185–193.

Maack, T. & Kersten, M. (2019). Finale Verluste von EU-Tochtergesellschaften: Der GAV als unüberwindbare Hürde für die grenzüberschreitende körperschaftsteuerliche Organschaft? *DStR*, 2281–2289.

Maul, S. (1998). Gerichtsstände und Vollstreckungsfragen bei konzernrechtlichen Ansprüchen gegenüber einem herrschenden Unternehmen im EG-Ausland. *AG*, 404–413.

Meilicke, W. (2002). Die Neuregelung der ertragsteuerlichen Organschaft über die Grenze. *DB*, 911–917.

Meilicke, W. & Sedemund, J. (2005). Welche Steuern sind nach dem Manninen-Urteil des EuGH in Deutschland anzurechnen? *DB*, 2040–2045.

Möhlenbrock, R. (2014). Niederlassungsfreiheit bei der Bildung steuerlicher Einheiten. *DB*, 1582–1583.

Micker, L. (2013). Die Aufgabe des doppelten Inlandsbezugs bei der Organschaft – Verlustverrechnung über die Grenze. *IWB*, 309–317.

Mitschke, W. (2010). Keine grenzüberschreitende Organschaft zum europarechtlichen „Nulltarif"! - Erwiderung auf den Beitrag von Klaus von Brocke in DStR 2010, 964 ff. *DStR*, 1368–1372.

Mitschke, W. (2011). Ergebnisabführungsvertrag „über die Grenze" und Abzug finaler Verluste ausländischer Tochtergesellschaften – Zugleich eine Erwiderung auf die Anmerkung von Homburg zu BFH-Urteil I R 16/10 (IStR 2011, 111). *IStR*, 185–189.

Müller, T., Detmering, M. & Lieber, B. (2020). *Die Organschaft – Körperschaftsteuerrecht, Gewerbesteuerrecht, Umsatzsteuerrecht, Grunderwerbsteuerrecht* (11. Aufl.). Herne.

Müller-Gatermann, G. (1997). Überlegungen zur Änderung der Organschaftsbesteuerung. In *Festschrift für Wolfgang Ritter* (S. 457–473). Köln: Kley, M. and Sünner, E. and Willemsen, A.

Mössner, J., Baumhoff, H. & Greif, M. (2018). *Steuerrecht international tätiger Unternehmen* (5. Aufl.). Köln.

Mössner, J., Seeger, S. & Oellerich, I. (2019). *Körperschaftsteuergesetz, Kommentar* (4. Aufl.). Herne.

Musil, A. & Weber-Grellet, H. (2018). *Europäisches Steuerrecht, Kommentar*. München.

Orth, M. (2005). Zurechnungsfragen bei Organschaft. *Der Konzern*, 79–97.

Pache, S. & Englert, M. (2007a). „Das Spiel ist aus!" – Kein positives Signal des EuGH für ein binnenmarktorientiertes Konzernbesteuerungsrecht. *IStR*, 844–850.

Pache, S. & Englert, M. (2007b). Grenzüberschreitende Verlustverrechnung deutscher Konzernspitzen - Ist die Organschaft noch zu retten? *IStR*, 47–53.

Picot, G. & Land, V. (1998). Der internationale Unternehmenskauf. *DB*, 1601–1607.

Prinz, U. (2000). Unternehmenssteuerreform 2001: Organschaftsbesteuerung im Wandel. *FR*, 1255–1262.

Prinz, U. (2010). Gedankensplitter zur konzeptionellen Fortentwicklung des steuerlichen Organschaftsrechts - Diskussionsbeitrag zum 2. Münchner Unternehmenssteuerforum. *Beihefter zu DStR*, 67–72.

Prinz, U. (2017). Finanzierungsbezogene Sonderbetriebsausgaben eines im Ausland ansässigen Mitunternehmers. *GmbHR*, 553–558.

Prinz, U. (2018). Verfehlte finanzierungsbezogene Abwehrgesetzgebung zu grenzüberschreitenden Mitunternehmerschaften. *DB*, 1615–1620.

Prinz, U. & Ludwig, F. (2020). Steuerliches Organschaftsrecht: Grenzüberschreitender Gewinnabführungsvertrag zu einer abhängigen EU/EWR-KapGes. mit inländischer Geschäftsleitung. *DB*, 1022–1029.

Prinz, U. & Witt, S. (2019). *Steuerliche Organschaft* (2. Aufl.). Köln.

Raupach, A. (2007). Das Steuerrecht – eine unerwünschte Quelle des Konzernrechts? In *Festschrift für Hans-Joachim Priester zum 70. Geburtstag* (S. 633–645). Köln: Hommelhoff, P. and Rawert, P. and Schmidt, K.

Raupach, A. & Pohl, D. (2005). Die Rechtssache Marks & Spencer - Konzernsteuerrecht auf dem Prüfstand des EuGH - Rückwirkungen auf das Konzernrecht? *NZG*, 489–492.

Rödder, T. (2012). Die kleine Organschaftsreform. *Ubg*, 717–723.

Rödder, T., Herlinghaus, A. & Neumann, R. (2015). *Körperschaftsteuergesetz, Kommentar*. Köln.

Rehfeld, L. (2014). Isolierte Gruppenbesteuerung zwischen Schwesterkapitalgesellschaften. *IWB*, 619–624.

Richelle, I., Schön, W. & Traversa, E. (2013). *Allocating taxing powers within*

the european union. Berlin.
Ritzer, C. & Aichberger, T. (2013). Die Anti-Double-Dip-Regelung des §14 Abs. 1 Satz 1 Nr. 5 KStG. *Der Konzern*, 602–614.
Rowedder, H. & Schmidt-Leithoff, C. (2017). *GmbHG - Gesetz betreffend die Gesellschaften mit beschränkter Haftung, Kommentar* (6. Aufl.). München.
Schaden, M. & Polatzky, R. (2013). Neuregelung der Verlustausgleichsbeschränkung des § 14 Abs. 1 Satz 1 Nr. 5 KStG – Auswirkungen auf deutsche Inbound-Finanzierungen über KG-Holding-Strukturen. *IStR*, 131–138.
Scheunemann, M. (2005). Europaweite Verlustberücksichtigung im Konzern: Steine statt Brot durch die Schlussanträge des Generalanwalts Maduro vom 7. 4. 2005 im Fall Marks & Spencer? *IStR*, 303–311.
Scheunemann, M. (2006). Praktische Anforderungen einer grenzüberschreitenden Verlustberücksichtigung im Konzern in Inbound- und Outboundfällen nach der Entscheidung Marks & Spencer. *IStR*, 145–154.
Scheunemann, M. (2015). *Grenzüberschreitende konsolidierte Konzernbesteuerung*. Köln.
Schiefer, F. (2014). (C-39/13, C-40/13, C-41/13 (Anm.)): Anmerkung zum Urteil des EuGH v. 12.06.2014, C-39/13, C-40/13, C-41/13. *EuZW*, 699–703.
Schirmer, H. (2013a). Holding als Organträger nach der Organschaftsreform. *GmbHR*, 797–799.
Schirmer, H. (2013b). Organschaft: Zuordnung zu einer inländischen Betriebsstätte. *FR*, 605–608.
Schneider, N. & Schmitz, S. (2013). Ausschluss der Verlustberücksichtigung bei Organschaft. *GmbHR*, 281–289.
Schönfeld, J. (2012). Praxisfragen der grenzüberschreitenden Organschaft – dargestellt anhand von Fallbeispielen. *IStR*, 368–374.
Schnitger, A. (2013). Fragestellungen zur steuerlichen Behandlung doppelt ansässiger Kapitalgesellschaften. *IStR*, 82–91.
Schnitger, A. (2014). Urteil des EuGH in der Rs. SCA als Katalysator für eine deutsche Organschaftsreform – jetzt geht's los (?). *IStR*, 587–591.
Schnitger, A. (2015). Keine 5 nichtabzugsfähige Betriebsausgaben in Folge grenzüberschreitender Organschaft – Auswirkungen der Entscheidung des EuGH in der Rs. Groupe Steria für den deutschen Rechtskreis. *IStR*, 772–774.
Schnitger, A. & Fehrenbacher, O. (2018). *Kommentar Körperschaftsteuer KStG, Kommentar* (2. Aufl.). Wiesbaden.
Schulze zur Wiesche, D. (2013). Die ertragsteuerliche Organschaft unter Berücksichtigung des Gesetzes zur Vereinfachung der Unternehmensbesteuerung und der steuerlichen Reisekosten sowie der aktuellen Rechtsprechung. *DStZ*, 621–630.
Schwenke, M. (2013). Grenzüberschreitende Organschaft – Anmerkungen zu den Neuregelungen im Gesetz zur Änderung und Vereinfachung der Unternehmensbesteuerung und des steuerlichen Reisekostenrechts. *ISR*, 41–46.
Sedemund, I., Jan/Sterner. (2006). Welche Folgen hat das Urteil „Marks & Spencer" für das deutsche internationale Steuerrecht? *DStZ*, 31–38.
Spahlinger, A. & Wegen, G. (2005). *Internationales Gesellschaftsrecht in der Praxis*. München.
Spindler, G. & Stilz, E. (2019). *Kommentar zum AktG, Kommentar* (4. Aufl.). München.
Stangl, I. & Brühl, M. (2013). Die „kleine Organschaftsreform". *Der Konzern*, 77–106.
Stöber, M. (2016). Die grenzüberschreitende Verlustverrechnung in Europa. *DStZ*, 582-592.
Streck, M. (2018). *Körperschaftsteuergesetz, Kommentar* (9. Aufl.). München.
Sydow, S. (2014). Gruppenbesteuerung: steuerliche Einheit zwischen Tochtergesellschaften – EuGH-Entscheidung in den verb. Rs. C-39 - 41/13 „SCA Group Holding BV". *IStR*, 480–486.
Thiel, J. (2004). Der fortschreitende Einfluss des EuGH auf die Ertragsbesteuerung der Unternehmen - Aktuelle Urteile und anhängige Verfahren. *DB*, 2603–2609.
Thoma, G. (2013). Unternehmensverträge i.S.d. §§291 ff. AktG mit nicht beteiligten Dritten – praktikables Instrument zur Konzernleitung oder strafbare Untreue? In *Festschrift für Michael Hoffmann-Becking zum 70. Geburtstag* (S. 1239–1259). Krieger, G. and Lutter, M. and Schmidt, K.
van Lishaut, I. (2009). Reform der Konzernbesteuerung (II). *FR*, 1030–1037.
van Venrooy, G. (1986). Isolierte Unternehmensverträge nach § 291 AktG? *BB*, 612–616.
Verlag C.H.BECK. (2018). Beck'sches Steuer- und Bilanzrechtslexikon, Loseblatt, München 2018 (Stand: Edition 50 v. 01.01.2020).
von Brocke, K. (2010). Abzug definitiver Verluste ausländischer Tochtergesellschaften im Rahmen der körperschaftsteuerlichen Organschaft? Zwei FG-Entscheidungen zur Anwendung der Grundsätze des EuGH in der Rs. Marks & Spencer. *DStR*, 964–968.
von Brocke, K. & Müller, S. (2014). Die Auswirkungen des SCA Group Holding-Urteils auf das deutsche Steuerrecht. *DStR*, 2106–2110.
von Brocke, K. & Müller, S. (2018). Europarecht. In *Konzernsteuerrecht* (3. Aufl., S. 66–92). München: Kessler, W. and Kröner, M. and Köhler, S.
von Freeden, A. & Liekenbrock, B. (2013). Neue Zinsabzugsbeschränkung für Inbound-Akquisitionsfinanzierungen durch § 14 Abs. 1 Nr. 5 KStG n. F.? *DB*, 1690–1696.
Wacker, R. (2017). Anmerkung zum Urteil des BFH v. 12.10.2016, I R 92/12. *IStR*, 278–288.
Wacker, R. (2018). Anmerkung zum Urteil des BFH v. 29.11.2017, I R 58/15. *FR*, 558–564.
Wagner, T. & Liekenbrock, B. (2013). Organschaft und Ausschluss der doppelten Verlustberücksichtigung im In- und Ausland nach § 14 Abs. 1 Nr. 5 KStG n.F. *Ubg*, 122–145.
Walter, W. (2005). Der Entwurf eines neuen BMF-Schreibens zur Organschaft. *GmbHR*, 456–459.
Walter, W. (2014). Gewinnabführungsvertrag mit Schwestergesellschaft aus zivilrechtlicher Sicht. *DB*, 2016–2017.
Walter, W. (2015). Organschaft zwischen Schwestergesellschaften de lege lata. *GmbHR*, 182–184.
Walter, W. (2017). Hat der BFH die Gewinngemeinschaft nun anerkannt? *Der Konzern*, 331-334.
Weigert, K. & Strohm, J. (2013). Zu den persönlichen Voraussetzungen der ertragsteuerlichen Organschaft unter Berücksichtigung aktueller Entwicklungen. *Der Konzern*, 249–272.
Weinberger, S. (2020). Der grenzüberschreitende Gewinnabführungsvertrag bleibt die große Hürde für die Anerkennung ausländischer Organgesellschaften. *Der Konzern*, 63–63.
Witt, C. (2009). Reform der Konzernbesteuerung (V). *FR*, 1045–2049.
Witt, S. (2010). Grenzüberschreitende Organschaft – Neue Entwicklungen der Rechtsprechung. *Ubg*, 737–745.
Wittkowski, A. (2008). *Grenzüberschreitende Verlustverrechnung in Deutschland und Europa – Eine ökonomische, europa- und verfassungsrechtliche Analyse*. Wiesbaden.

Junior Management Science 6(3) (2021) 590-636

Junior Management Science

journal homepage: www.jums.academy

Backtesting the Expected Shortfall

Konstantin Spring

Universität Konstanz

Abstract

Backtesting of risk measure estimates is an integral part for an effective risk management. With the growing importance of the Expected Shortfall (ES) to potentially replace the Value at Risk (VaR) as a primary measure for market risk this also calls for suitable backtesting solutions. Although a variety of approaches has been proposed in the past, there is still an on-going discussion whether the ES can be properly backtested. The thesis adds to this discussion in the following way. Five of the most promising backtests for the ES are implemented, compared based on theoretical properties like empirical size and power and tested against ES estimation models which are fitted to historical returns of the S&P 500. In addition, all backtests in scope are assessed against a set of criteria which reflect their practical applicability for both regulators and financial institutions. Results presented within this thesis confirm that backtesting the ES is indeed not much more complicated than backtesting the VaR. Backtesting ES might be conceptually less straight forward, but there are multiple promising approaches which allow for a reasonable validation of ES estimation models.

Keywords: Expected shortfall; backtesting; risk measures; statistical test.

1. Introduction

This thesis deals with the ability to backtest the *Expected Shortfall (ES)* risk measure. Indeed, backtesting the ES has been one of the most interesting topics in the field of financial risk management over the last two decades. Until today there is no consensus amongst researches on how to backtest the ES. Furthermore, there is still an ongoing and controversial debate on whether it is possible to backtest the ES at all. This thesis aims to shed light on the diverse literature on backtesting the ES. Moreover, it yields a comprehensive evaluation of multiple backtesting methodologies, suggested within the literature, also with regard to their practical applicability for financial institutions.

Within the financial industry, it is common practice to make use of so called *risk measures* in order to determine potential risks inherited in an investment portfolio. Risk measures are mappings of random variables, i.e. the *Profit and Loss (P&L) distribution* of a portfolio, into real-valued scalars, which represent the capital amount that needs to be held by a financial institution as a buffer against unexpected portfolio losses. Amongst practitioners, the ES together with the *Value at Risk (VaR)* are by far the two most relevant risk measures, which will both be defined in more detail in the course of this thesis.

Besides the estimation of these risk measures, it is fundamental, especially for regulators, to be able to ex-ante verify the accuracy of a risk forecasting model, to ensure that capital buffers are sufficient. This process is referred to as the *backtesting* of a certain risk measure. As outlined by Kratz, Lok, and McNeil (2018), a *backtest* is a statistical procedure which compares forecasts to actual realizations in order to judge the correctness of the implemented forecasting model. In the case of risk management, a set of ex-ante VaR or ES estimates is compared to realized portfolio returns in order to verify if the applied estimation model accurately forecasts the risk of the underlying portfolio. For regulators it is of particular interest to detect estimation models which underestimate the actual underlying risk, as an insufficient coverage against adverse scenarios might threaten the functioning of the whole financial system in times of market distress.

For many years the VaR has been the predominant risk measure in practice. Nevertheless, it is well known that the VaR does not belong to the class of *coherent* risk measures, defined by Artzner, Delbaen, Eber, and Heath (1999), as it fails to consistently account for portfolio diversification. Furthermore, the VaR only accounts for a predefined quantile of the P&L distri-

DOI: https://doi.org/10.5282/jums/v6i3pp590-636

bution but ignores the tail of extreme events beyond that quantile, which is for example noted in Danielsson et al. (2001). Thus, it appears to be consequential, that the Basel Committee of Banking Supervision in their "Fundamental Review of the Trading Book" (see Basel Committee (n.d.-a)), proposes to replace the VaR as a primary measure for market risk by the ES. Indeed, the ES makes up for the weaknesses of the VaR listed above, which will be outlined in the course of the following chapter. Furthermore, as for example argued in Emmer, Kratz, and Tasche (2015), most forecasting models for the VaR can easily be generalised for the estimation of the ES.

Nevertheless, there is one major concern about the ES both practitioners and researchers worry about - the ability to backtest the ES. Indeed, backtesting VaR is more or less intuitive and the respective literature is well developed.[1]. In comparison to the VaR, there is still an ongoing discussion on whether the ES is even backtestable at all. This was originally fueled by Gneiting (n.d.), who proved that the ES lacks a statistical property known as *elicitabilty*. As shortly summarized by Kratz et al. (2018), a risk measure is called elicitable, if it can be defined as the solution of a forecasting-error minimization problem. Further, Gneiting (n.d.) argues that the lack of this property makes it cumbersome or maybe even impossible to backtest the ES. This opinion is further supported by contributions like Chen (2014) or Carver (2013). Nevertheless another opinion starts to prevail in recent years. As an example, Acerbi and Szekely (2014) makes the point, that elicitability has actually nothing to do with backtesting at all. The concept of elicitability as well as any potential implications on the ability to backtest the ES will also be taken up in the course of this thesis.

Nevertheless and despite all doubts expressed within the literature, a variety of potential backtests for the ES has been suggested in the last two decades[2]. This thesis tries to provide insights into the diverse and still emerging literature by implementing some of the most promising approaches. More precisely, five different approaches are selected, based on Kratz et al. (2018), Acerbi and Szekely (2014), Bayer and Dimitriadis (2019), Costanzino and Curran (2015), and McNeil and Frey (2000), which are evaluated in the course of this thesis. Furthermore, some adjustments to the original test version are proposed, in order to ensure both practical applicability as well as compliance with regulatory needs. All backtests are implemented in *Python* and compared according to classical measures like empirical power and size. Additionally, they are tested in the their judgement of actual forecasting models, which are fitted to financial returns of the S&P 500.

The remainder of this Master thesis is structured as follows. Chapter 2 is going to introduce the concept of risk measures, define both the VaR and the ES and outline their theoretical properties. Afterwards, chapter 3 focuses on the concept of backtesting in general, whereas chapter 4 introduces the selected approaches as well as adjustments, which will be made for the purpose of this thesis. Consecutively, chapter 5 conducts a simulation study to evaluate and compare empirical size and power of the applied tests, followed by an application of all backtests to real financial data in chapter 6. Finally, some concluding remarks on the results obtained within this thesis are presented in chapter 7.

[1] See for example Kupiec (1995) or Christoffersen (1998) for work on backtesting VaR.

[2] See for example Acerbi and Szekely (2014), Bayer and Dimitriadis (2019), Costanzino and Curran (2015), Du and Escanciano (2017), Kratz et al. (2018), Wong (2008), Berkowitz (2001), Nolde and Ziegel (2017), McNeil and Frey (2000) and many more for approaches on backtesting the ES.

2. Risk Measures and their properties

The purpose of the following section is twofold. First, it gives an overview on the concept of risk measures and introduces some notation used throughout this thesis. Based on Artzner et al. (1999), Emmer et al. (2015) and Nolde and Ziegel (2017), desirable properties of risk measures are outlined. Second, the two most practically relevant risk measures, i.e. the Value at Risk (VaR) and the Expected Shortfall (ES) are formally defined. Furthermore, the section aims to give a theoretical foundation for the decision by regulators to replace the VaR by the ES as a primary measure for market risk. Thus, this section might be seen as a motivation why it is even necessary to pay attention to the backtestability of the ES.

2.1. Concept of Risk Measures

For the general definition of a risk measure, Artzner et al. (1999) is taken as a reference. Therefore, a position, e.g. an investment strategy or a portfolio of financial assets, is considered over a certain time horizon. The position might be described by X, its net worth at the end of the investment horizon, which is defined as a real-valued function on some set Ω of possible future outcomes. Moreover, let $\mathscr{X}$ be the set of all real-valued functions $X \in \mathscr{X}$, which represent the future net worth of a certain position. The formal definition of a risk measure is then given as follows.

Definition 2.1 (Risk Measure)**.** *A risk measure ρ is a mapping from $\mathscr{X}$ to $\mathbb{R}$, i.e.*

$$\rho : \mathscr{X} \to \mathbb{R}, \quad X \mapsto \rho(X). \tag{2.1}$$

Artzner et al. (1999) connects the concept of risk measurement to the particular question how close or far a certain position is from being accepted by either regulators or other stakeholders. More general speaking, $\rho(X)$, which is related to a certain end of period net worth $X \in \mathscr{X}$, is the additional capital amount that needs to be held such that the position is accepted by regulators. Correspondingly, a higher value of $\rho(X)$ is related to a riskier investment position. For instance, a value of $\rho(X) > 0$ signals that a financial institution needs to hold additional capital as a safety buffer such that the position is accepted by regulators. On the other hand, a position is accepted by regulators if $\rho(X) \leq 0$. As a natural extension, Artzner et al. (1999) thus defines the so called *acceptance set* $\mathscr{A}_\rho$, which contains all positions acceptable for regulators with respect to a certain risk measure.

Definition 2.2 (Acceptance Set)**.** *The acceptance set related to a risk measure ρ is denoted by $\mathscr{A}_\rho$ and defined through*

$$\mathscr{A}_\rho := \{X \in \mathscr{X} | \rho(X) \leq 0\} \tag{2.2}$$

Overall, the concept of risk measures facilitates the decision making of both financial institutions and regulators as it breaks down the risks inherited in a certain position to one single number. Furthermore, risk measure values ought to be easy to interpret as the amount of additional capital that needs to be held by an institution.
Obviously, a mapping ρ from $\mathscr{X}$ into the real values should fulfil some desirable properties to be an adequate risk measure for practical needs. Nevertheless, the following subsection is first going to introduces both the VaR as well as the ES, before the subsection thereafter discusses their theoretical properties.

2.2. Value at Risk, Expected Shortfall and Spectral Risk Measures

There are several different conventions which are used throughout the literature for the definition of both the VaR as well as the ES. In order to maintain consistency, the following set-up is going to be used in the remainder of this thesis.
For the previous general introduction to risk measures, the future net worth of a position, X, was the variable of interest. Nevertheless, practitioners rather consider the Profit & Loss (P&L) distribution of a certain position. Moreover, for the case of risk management especially the estimation of a loss distribution is of particular interest. Consequently, I decided to rely on the following set of notations.

Notation 2.3. *The following notation is going to be used in the course of the thesis.*

- *Denote the unconditional return loss L of a portfolio, as a real valued random variable on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$. Furthermore, let $F(y) := \mathbb{P}(L \leq y)$ be the cumulative distribution function (CDF) of L and denote the respective probability density function (PDF) by $f(y)$.*
- *Denote the conditional return loss $\{L_t\}_{t\geq 0}$ of a portfolio as a real valued stochastic process on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$, endowed with a filtration $\mathscr{F} := \{\mathscr{F}_t, t \geq 0\}$. The filtration is given by $\mathscr{F}_t := \sigma\{L_s, s \leq t\}$, for all $t \in \mathbb{N}$.*
- *For $t \in \mathbb{N}$, denote the conditional CDF of L_t based on the set of past information $\mathscr{F}_{t-1}$ by $F_t(y) := \mathbb{P}(L_t \leq y|\mathscr{F}_{t-1})$. Respectively, denote the corresponding conditional PDF by f_t.*
- *For $t \in \mathbb{N}$, denote the conditional quantile of L_t given the past information $\mathscr{F}_{t-1}$ and some level $\alpha \in (0,1)$ by $q_\alpha(L_t|\mathscr{F}_{t-1}) := F_t^{\leftarrow}(\alpha) = \inf\{y \in \mathbb{R} : F_t(y) \geq \alpha\}$, where $F_t^{\leftarrow}$ denotes the generalized inverse function of F_t.*

A summary of the most important variables used throughout this thesis can be found in Appendix B attached to this thesis.

In Notation 2.3, L is defined as the random variable of return losses, nevertheless in some publications L is prescribed to be the distribution of return losses, which does make a difference form a mathematical point of view. As noted by Nolde and Ziegel (2017), defining a risk measure ρ on a space of return loss distributions instead of a space of random variables, only makes sense if the risk measure ρ is law invariant. This means that two equivalent return loss distributions, i.e. $F^1(y) = F^2(y)$ for all $y \in \mathbb{R}$ and for some random variables L^1 and L^2, always results in the same risk measure value, i.e. $\rho(L^1) = \rho(L^2)$. Indeed, both the VaR and the ES satisfy the criteria of law-invariance, thus it is not crucial for the purpose of this thesis whether L is defined as a random variable or as the related return loss distribution. Overall, I want to rely on the Notation 2.3, but I might at certain points mix up the notation, if this fits better into the commonly applied notion of a certain topic.
Especially in the area of market risk, it is well known that the financial environment and thus the risk inherited in a investment portfolio changes over time. Thus, modelling the unconditional loss L of a portfolio can be seen as a rather naive approach. Consequently, the main focus of this thesis will be on conditional, time dependent, risk forecasts based on the estimated stochastic process of $\{L_t\}_{t\geq 0}$.
From a historical perspective both variance and standard deviation were the dominant risk measures in financial applications for a long period of time. The importance of the VaR started to raise when RiskMetrics announced to use the VaR as their standard measure of risk in 1996, as outlined in JP Morgan (1996). Shortly thereafter, the Basel Committee of Banking Supervision introduced the VaR as an industry standard for the measurement of market risk (see Basel Committee (n.d.-c)). In order to calculate the VaR one needs to consider a time horizon Δ as well as a confidence level $\alpha \in (0,1)$.[3] Given a return loss variable L_t, the VaR describes the loss over the considered time horizon Δ, which is only exceeded in $1-\alpha$ percent of all cases. For the remainder of this thesis, a risk horizon of $\Delta = 1$ day is considered and thus Δ can be dropped in the following definition in order to simplify the notation.

Definition 2.4 (Value at Risk)**.** *The Value at Risk (VaR) at time $t \geq 0$, $t \in \mathbb{N}$, at the confidence level $\alpha \in (0,1)$ based on some conditional return loss L_t is defined as*

$$VaR_\alpha(L_t|\mathscr{F}_{t-1}) := q_\alpha(L_t|\mathscr{F}_{t-1}) = \inf\{y \in \mathbb{R} : F_t(y) \geq \alpha\} \,. \quad (2.3)$$

For simplicity reasons, the short notation $VaR_{t,\alpha} = VaR_\alpha(L_t|\mathscr{F}_{t-1})$ is used.

Analogous to Definition 2.4, one can define a time independent version VaR_α based on unconditional return loss L. The Basel Committee of Banking Supervision requires executives to calculate the VaR at a confidence level of $\alpha = 0.99$. Given any realistic portfolio and the conventions applied in Definition 2.4, it is thus reasonable to assume that the $VaR_{t,0.99}$ has a positive sign. The VaR is often criticized for two main reasons. First of all, it is insensitive to any extreme losses beyond the α-quantile of the loss distribution. Secondly, it is not a sub-additive risk measure, i.e. it fails to consistently account for portfolio diversification. This will be outlined in more depth in subsection 2.3.

[3] Note that α is selected as a confidence level and not a significance level for the purpose of this thesis. Thus, α takes on values like 0.95 or 0.99 instead of 0.05 or 0.01.

[4]

As already shortly mentioned in the introduction, the ES overcomes both of these shortcomings and thus gains more and more attention in the measurement of market risk, as for example the Basel Committee of Banking Supervision suggests to replace the VaR by the ES under the Basel III framework[5]. Indeed, the ES appears to be a natural extension of the VaR. Whereas the VaR states the loss which is only exceeded at a certain confidence level α, the ES goes one step further and asks for the average loss given an exceedance of the corresponding VaR threshold. More formally, given the notations introduced above, the ES is defined as follows.

Definition 2.5 (Expected Shortfall)**.** *The Expected Shortfall (ES) at time $t \geq 0$, $t \in \mathbb{N}$, at the confidence level $\alpha \in (0,1)$ based on some conditional return loss L_t with CDF F_t is defined as*

$$\begin{aligned} ES_\alpha(L_t|\mathscr{F}_{t-1}) &:= \frac{1}{1-\alpha}\int_\alpha^1 q_p(L_t|\mathscr{F}_{t-1})dp \qquad (2.4)\\ &= \frac{1}{1-\alpha}\int_\alpha^1 VaR_{t,p}dp\,. \end{aligned}$$

Moreover, if the CDF F_t is continuous at the α-quantile, the definition can be simplified, according to Acerbi and Tasche (2002), to

$$ES_\alpha(L_t|\mathscr{F}_{t-1}) = \mathbb{E}[L_t|L_t \geq VaR_{t,\alpha}]\,. \qquad (2.5)$$

Again the short notation $ES_{t,\alpha} = ES_\alpha(L_t|\mathscr{F}_{t-1})$ will be applied.

The value of the ES again depends on the arbitrary confidence level $\alpha \in (0,1)$. As the conditional quantile $q_\alpha(L_t|\mathscr{F}_{t-1})$ is an increasing function in α, it holds that $ES_{t,\alpha} \geq VaR_{t,\alpha}$. Kerkhof and Melenberg (2004) argues that the overall level of capital requirements, applying $VaR_{t,0.99}$, is more or less appropriate. Thus, in order to keep capital requirements on a similar level one should choose a lower confidence level for the ES compared to the VaR. Indeed, Basel Committee (n.d.-b) suggests the estimation of the ES at a level of $\alpha = 0.975$. If one assumes that the return loss follows a standard normal distribution, than both $VaR_{t,0.99}$ as well as $ES_{t,0.975}$ lead to similar values. Following Kerkhof and Melenberg (2004), one might denote the PDF and the CDF of the standard normal distribution by ϕ and Φ, respectively. Then it holds [6],

$$\begin{aligned} VaR_{t,0.99} &= \Phi^{-1}(0.99) = 2.33 \approx 2.34 \qquad (2.6)\\ &= \frac{\phi(1.96)}{0.025}\\ &= \frac{\phi(\Phi^{-1}(0.975))}{1-0.975} = ES_{t,0.975}. \end{aligned}$$

Therefore, the overall magnitude of the capital buffer remains comparable. Nevertheless, if the loss distribution of a portfolio exhibits excess kurtosis compared to the standard normal distribution, this leads to a situation with $ES_{t,0.975} > VaR_{t,0.99}$. Consequently, the $ES_{t,0.975}$ measure captures the additional risk inherited in a heavy tailed loss distributions. Therefore, the ES is more sensitive to potential extreme losses in the underlying portfolio, which is beneficial from a risk management perspective. For the remainder of this thesis, the ES will always be considered at the confidence level of $\alpha = 0.975$, stipulated by the Basel Committee.

For a further classification of both the VaR and the ES the term of a *spectral risk measure*, which stems from Acerbi (2002), is introduced below. As for example described in Costanzino and Curran (2015), a spectral risk measure can be seen as a weighting of the VaR at different confidence levels given some spectrum ψ. According to Costanzino and Curran (2015), a so called *admissible risk spectrum* needs to fulfil the following properties.

Definition 2.6 (Admissible risk spectrum)**.** *An integrable function $\psi \in L^1([0,1])$ is called an admissible risk spectrum, if*

(i) *ψ is non-negative on $[0,1]$,*

(ii) *ψ is non-decreasing on $[0,1]$,*

(iii) *$\|\psi\|_1 = 1$, where $\|\psi\|_1 = \int_0^1 |\psi(p)|dp$.*

Based on Definition 2.6, a spectral risk measure can be defined in the following way.

Definition 2.7 (Spectral risk measure)**.** *Let L_t be a conditional return loss with some cumulative distribution function F_t. Suppose the ψ is an admissible risk spectrum, then the spectral risk measure $\mathscr{M}_\psi$ based on the risk spectrum ψ, at time $t > 0$, $t \in \mathbb{N}$, is defined by*

$$\mathscr{M}_{t,\psi} := \int_0^1 \psi(p) VaR_{t,p}dp. \qquad (2.7)$$

Given Definition 2.7, one can easily recognize that the VaR has no representation as a spectral measure, whereas the ES does belong to that class of risk measures. This is explained in the consecutive Lemma, following Costanzino and Curran (2015).

Lemma 2.8. *As outlined in Costanzino and Curran (2015) it holds:*

(i) *The VaR defined in Definition 2.4 is not a spectral risk measure.*

(ii) *The ES defined in Definition 2.5 is a spectral risk measure.*

Proof. (i) Define $\psi_{VaR}(p) := \delta_{\alpha,p}$, where δ denotes the Kronecker-Delta, then it holds

$$\mathscr{M}_{t,\psi_{VaR}} = \int_0^1 \delta_{\alpha,p} VaR_{t,p}dp = VaR_{t,\alpha}\,. \qquad (2.8)$$

But ψ_{VaR} is not an admissible risk spectrum, as it violates conditions (ii) and (iii) from Definition 2.6.

[4] Note that I defined L_t as the return loss and not as an absolute loss figure. Consequently, the resulting VaR is also on a return space, i.e. it needs to be multiplied with the corresponding position size to be interpreted as an absolute value in a certain currency.

[5] See for example Basel Committee (n.d.-a) and Basel Committee (n.d.-b).

[6] See chapter 6, formulas (6.2) and (6.3) for the calculation of VaR and ES under the assumption of normally distributed return losses.

(ii) Define $\psi_{ES}(p) := \frac{1}{1-\alpha}\mathbb{1}_{\{\alpha\leq p\leq 1\}}$, then it holds

$$\begin{aligned}\mathcal{M}_{t,\psi_{ES}} &= \int_0^1 \frac{1}{1-\alpha}\mathbb{1}_{\{\alpha\leq p\leq 1\}} VaR_{t,p}\,dp \qquad (2.9)\\ &= \frac{1}{1-\alpha}\int_\alpha^1 VaR_{t,p}\,dp\\ &= ES_{t,\alpha}.\end{aligned}$$

Furthermore, ψ_{ES} fulfils all three properties of an admissible risk spectrum as defined in Definition 2.6.

□

Based on Lemma 2.8, one can rely on general results on the class of spectral risk measures to differentiate between the properties fulfilled by both the VaR and the ES.

2.3. Basic properties of Risk Measures

This subsection aims to give an overview on basic properties of risk measures, which are widely accepted throughout the literature. Furthermore, both the VaR and the ES are going to be evaluated based on the introduced properties.

As previously mentioned, Artzner et al. (1999) brought up the term of a *coherent risk measure*, which needs to fulfil a set of mathematical axioms. The notation for the following definition is based on Emmer et al. (2015)[7].

Definition 2.9 (Coherent Risk Measure)**.** *A risk measure ρ is called a coherent risk measure, if it satisfies the following properties:*

(i) Homogeneity:
A risk measure ρ is homogeneous, if for any return loss variable L and $h\in\mathbb{R}$, $h\geq 0$ it holds that:

$$\rho(hL) = h\rho(L)\,. \qquad (2.10)$$

(ii) Monotonicity:
A risk measure ρ is monotonic, if for any two return loss variables L^1 and L^2 it holds that:

$$L^1\leq L^2 \Rightarrow \rho(L^1)\leq\rho(L^2)\,. \qquad (2.11)$$

(iii) Translation invariance:
A risk measure ρ is translation invariant, if for any return loss variable L and for any $m\in\mathbb{R}$ it holds that:

$$\rho(L-m) = \rho(L)-m\,. \qquad (2.12)$$

(iv) Sub-additivity:
A risk measure ρ is sub-additive, if for any return loss variables L^1 and L^2 it holds that:

$$\rho(L^1+L^2)\leq\rho(L^1)+\rho(L^2) \qquad (2.13)$$

[7] Again different conventions are used for the definition of a coherent risk measure. Nevertheless, the notation used by Emmer et al. (2015) fits best into the previously introduced notation.

Indeed, all four axioms given in Definition 2.9 are reasonable properties risk measures should satisfy. Homogeneity states that the risk inherited in a portfolio should be scalable by the respective size of the portfolio.[8] Secondly, if a portfolio $\mathbb{P}$-almost surely generates a higher loss than another portfolio, i.e. $L^2\geq L^1$, then that portfolio should also be related with a higher riskiness, i.e. $\rho(L^2)\geq\rho(L^1)$, which is fulfilled if the risk measure is monotonic. Furthermore, if the portfolio loss is reduced by adding a risk-free capital amount m to the portfolio, then the risk should decrease by exactly the amount of m, which is satisfied if the respective risk measure is translation invariant. Lastly, the risk inherited in one large portfolio consisting of multiple positions, should be at most as large as summing up the risks of the single positions due to diversification effects. This is fulfilled whenever a risk measure ρ is sub-additive.

Acerbi (2002) even goes one step further and includes conditions (i)-(iv) in its general definition of a risk measure. Moreover, he argues that these conditions should be fundamental for every risk measure. For the particular case of both the VaR and the ES, Acerbi (2002) derives an essential result, which relates both the class of spectral and coherent risk measures.

Theorem 2.10 (Acerbi (2002))**.** *Let L_t be some conditional return loss variable with CDF F_t. Define $\mathcal{M}_{t,\psi}$ as*

$$\mathcal{M}_{t,\psi} := \int_0^1 \psi(p)VaR_{t,p}\,dp,$$

where $\psi\in L^1([0,1])$. Then $\mathcal{M}_{t,\psi}$ is a coherent risk measure if and only if ψ is an admissible risk spectrum according to Definition 2.6.
Moreover, a risk measure is coherent if and only if it has a representation as a spectral risk measure.

Proof. See Theorem 2.5 and Theorem 4.1 in Acerbi (2002). □

As the VaR does not have a spectral representation, it is also not a coherent risk measure following Theorem 2.10. On the contrary, the ES does belong to the class of coherent risk measures. More precisely, the VaR fails to satisfy condition (iv) of a coherent risk measure, i.e. it is not sub-additive. One can construct counter-examples with $n\in\mathbb{N}$ loss distributions $L^1,\ldots,L^n$ such that $\rho\left(\sum_{i=1}^n L^i\right) > \sum_{i=1}^n \rho(L^i)$.[9] The lack of sub-additivity might be seen as a drawback of the VaR risk measure in practical applications. Indeed, portfolio diversification is a common method of reducing financial risks. Nevertheless, if the achieved diversification effects are not consistently reflected in the related VaR figures, this might yield wrong incentives for financial institutions.

Emmer et al. (2015) also list another property which appears to be essential for risk measures in practical applications. The *comonotonic additivity* of a risk measure can be seen as a complement to the sub-additivity condition stated in Definition 2.9. The property is based on the following two definitions.

[8] Note that in the presents of liquidity risk for large position sizes one might rather expect $\rho(hL)\geq h\rho(L)$ given that $h\gg 0$.

[9] See Hull (2015) Examples 12.5 and 12.6, for examples on the VaR where the sub-additivity condition is violated.

Definition 2.11 (Comonotonicity)**.** *Let X^1 and X^2 be two real valued random variables on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$. Then X^1 and X^2 are called comonotonic, if there exists a real-valued random variable $\tilde{X}$ and two non-decreasing functions f^1 and f^2, such that*

$$X^1 = f^1(\tilde{X}) \text{ and } X^2 = f^2(\tilde{X}) \,. \tag{2.14}$$

Definition 2.12 (Comonotonic Additivity)**.** *Let L^1 and L^2 be two comonotonic return loss variables, then a risk measure ρ is called comonotonic additive, if it holds that*

$$\rho(L^1 + L^2) = \rho(L^1) + \rho(L^2) \,. \tag{2.15}$$

Indeed, this property has also a comprehensible intuition. If the loss distributions of two portfolios are comonotonic, then their losses are both driven by some common risk factor. Thus, merging both portfolios together should not lead to any diversification benefits as the risk stems from the same risk factor anyway. Actually, both the VaR as well as the ES satisfy the property of comonotonic additivity.[10]

Besides the above listed properties, the aspect of *robustness* is also often discussed in the context of risk measures. In a general sense, a risk measure is said to be robust if small changes in the underlying loss distribution are also related to small changes in the risk measure value. More formal definitions of robustness can for example be found in Cont, Deguest, and Scandolo (2008) or Emmer et al. (2015). Cont et al. (2008) evaluates how additional data points, added to the estimation sample for a return loss variable, effect overall risk measure values. Indeed, they conclude that robustness is not solely a matter of the selected risk measure, but does also depend on the type of implemented estimation model. Furthermore, they outline that the ES is in general less robust compared to the VaR. Nevertheless, as for example noted by Emmer et al. (2015), this is not overly surprising as the ES was introduced due to the insensitivity of the VaR with respect to observations beyond the α-quantile. Thus, it is by definition more or less a logical implication that the ES is less robust compared to the VaR. Furthermore, Emmer et al. (2015) argues that in the context of risk management, extreme observations might be less related to measurement errors but rather to actual circumstances, which should indeed be reflected by the respective risk measure. Summing up, the ES is obviously less robust compared to the VaR, at least in the sense of Cont et al. (2008), but it is not clear whether this should be seen as a potential shortcoming of the ES, given that a certain degree of sensitivity towards extreme observations lies within the nature of a risk measure.

The properties presented in this section display just a small selection. Overall, there is far more literature on general properties of risk measures.[11] Nevertheless, as a result of this subsection, it can be seen that the VaR is not a coherent risk measure as it does not fulfil the axiom of sub-additivity. Furthermore, the VaR is only defined as a quantile of the return loss distribution, consequential it completely ignores all losses beyond the selected confidence level.

Concluding this subsection, it appears to be reasonable to favour the ES over the VaR, given that it is a coherent risk measure which furthermore satisfies the property of comonotonic additivity. In addition, also the lack of robustness should, at least per-se, not be seen as a disadvantage of the ES, given the previous argumentation. Nevertheless, for a long time the majority of both academics as well as practitioners expressed their reservations with regard to the ES, as they doubted the ability to backtest the ES. Just in recent time, triggered by the decision of the Basel Committee of Banking Supervision in 2013, researchers started to focus more on practical backtesting solutions instead of pointing towards potential conceptual limitations.

2.4. Elicitability

As shortly described in the introduction, the statistical property of elicitability is always closely related to any discussion on the backtestability of the ES. Thus, this subsection gives a brief summary on the elicitability of risk measures or functionals in a more general context.

Originally, the concept of elicitiability was introduced by Osband (1985). Overall, as for example stated by Gneiting (n.d.), elicitability is a useful tool for the evaluation of a series of forecasts based on ex-ante realizations. The concept makes use of a *strictly consistent scoring function S*, which needs to satisfy a set of criteria, and takes both forecasts and realizations as input parameters in order to assign them to a numerical score. A functional, i.e. a risk measure in this case, is said to be elicitable, if a suitable strictly consistent scoring function exists, which fulfils the properties listed in the course of this subsection.

Acerbi and Szekely (2014) briefly explains the main advantage of an elicitable risk measures. Let ρ be some elicitable risk measure with strictly consistent scoring function S. Assume a risk manager estimates some return loss variable $\widehat{L}$, which results in a risk forecast $\rho(\widehat{L})$. Furthermore, let L denote the true distribution of return losses. Then the true value of the risk measure, $\rho(L)$, can be determined by

$$\rho(L) = \arg\min_{\rho(\widehat{L})} \mathbb{E}[S(\rho(\widehat{L}), L)]. \tag{2.16}$$

In practice, a risk manager generates a set of forecasts $\{x_t = \rho(\widehat{L}) : t = 1, \ldots, T\}$ and observes a set of realizations $\{l_t, t = 1, \ldots, T\}$, which are based on the true data generating process L. Given both forecasts and realizations, one might try to minimize the mean-score,

$$\bar{S} = \frac{1}{T} \sum_{t=1}^{T} S(x_t, l_t), \tag{2.17}$$

in order to find a forecasting distribution $\widehat{L}$ such that the achieved risk estimates are close to the true values, i.e. $\rho(\widehat{L}) \approx \rho(L)$.

In the following a slightly more formal definition of elicitability is presented, based on Emmer et al. (2015) and Nolde and Ziegel (2017). For a more comprehensive and also technical

[10] See for example Emmer et al. (2015) for the definition of Expectiles, a coherent risk measure which does not satisfy comonotonic additivity.

[11] See for example Foellmer and Schied (2002) and its definition of a convex risk measure for another important contribution on risk measure properties.

review on the concept of elicitability see for example Gneiting (n.d.) or Fissler and Ziegel (2016). In line with common practice, L denotes the distribution of return losses and not a random variable for the purpose of this definition. In general, the elicitability of a risk measure always depends on some set of probability distributions, i.e. loss distributions in the setting of this thesis. Denote the set of considered probability distributions by $\mathscr{P}$. Furthermore, similar as in Nolde and Ziegel (2017), let $\Theta(L) = (\rho^1(L), \dots, \rho^k(L))$ be a vector of $k \geq 1$, $k \in \mathbb{N}$ risk measures based on some return loss distribution $L \in \mathscr{P}$.
The following formal definitions of a strictly consistent scoring function an the elicitability of risk measures follow Nolde and Ziegel (2017). Note that for the case of $k = 1$, i.e. $\Theta(L) = \rho^1(L)$ the elicitability of one single risk measure is evaluated, while for the case of $k > 1$ the joint elicitability of a vector of risk measures is considered.

Definition 2.13 (Consistency)**.** *A scoring function $S : \mathbb{R}^k \times \mathbb{R} \to [0, \infty)$ is called*

(i) consistent for Θ with respect to $\mathscr{P}$ if

$$\mathbb{E}[S(\Theta(L), L)] \leq \mathbb{E}[S(x, L)] \,, \tag{2.18}$$

for all $x \in \mathbb{R}^k$ with $x = (x^1, \dots, x^k) \neq \Theta(L) = (\rho^1(L), \dots, \rho^k(L))$ and for all loss distributions $L \in \mathscr{P}$.

(ii) strictly consistent for Θ with respect to $\mathscr{P}$ if equation (2.18) holds with strict inequality, i.e.

$$\mathbb{E}[S(\Theta(L), L)] < \mathbb{E}[S(x, L)] \,, \tag{2.19}$$

for all $x \in \mathbb{R}^k$ with $x = (x^1, \dots, x^k) \neq \Theta(L) = (\rho^1(L), \dots, \rho^k(L))$ and for all loss distributions $L \in \mathscr{P}$.

Definition 2.14 (Elicitability)**.** *A vector Θ of risk measures is called elicitable with respect to $\mathscr{P}$, if there exists a strictly consistent scoring function $S : \mathbb{R}^k \times \mathbb{R} \to [0, \infty)$ in the sense of Definition 2.13.*

As outlined above, elicitability can be seen as a helpful tool for the purpose of risk management as it yields a criterion for the estimation of an optimal forecast. Indeed, given a strictly consistent scoring function, different estimated loss distributions and the resulting risk forecasts can be compared with respect to the minimization problem stated in formula (2.16).
In the univariate case, i.e. for $k = 1$, the most prominent example might be the mean functional, which is elicitable with respect to all real valued probability distributions with finite second moment. Moreover, the squared error, i.e. $S(x, y) = (x - y)^2$, is a strictly consistent scoring function in this case. Thus, for example the estimation of a mean regression in a standard OLS framework, is achieved by minimizing the mean-squared-error, which corresponds to minimizing formula (2.17) in this particular case.
In order to remain within the scope of this thesis, only the major results regarding the elicitability of different risk measures are cited below. The VaR at confidence level α is elicitable with respect to the class of real-valued Borel-probability distributions, which have a unique α-quantile. Moreover, the weighted absolute error function is a strictly consistent scoring function for the VaR, as shown in Thomson (1979) and Saerens (2000). On the contrary, as previously mentioned, Gneiting (n.d.) proved that the ES is not an elicitable risk measure. In more detail, the following proposition was stated by Gneiting (n.d.).

Proposition 2.15 (Gneiting (n.d.), Theorem 11)**.** *The ES functional is not elicitable relative to any class $\mathscr{P}$ of real-valued probability distributions on the interval $I \subset \mathbb{R}$ that contains the measures with finite support, or the finite mixtures of the absolutely continuous distributions with compact support.*

Thus, whereas the VaR is elicitbable with respect to a reasonable set of probability distributions, the ES, at least standalone, is not.
Nevertheless, for the case of $k = 2$, recent contributions derived, that the vector $\tilde{\Theta} := (VaR_\alpha, ES_\alpha)$ is indeed elicitable following Definition 2.14. In the literature, this is often referred to as the conditional elicitability of the pair of ES and VaR.
The following result, proved by Fissler and Ziegel (2016), is cited below as is fundamental for the backtest by Bayer and Dimitriadis (2019), which will be introduced in the course of this thesis.

Proposition 2.16 (Nolde and Ziegel (2017))**.** *All scoring functions of the form,*

$$\begin{aligned} S(x_1, x_2, y) \quad &= \mathbb{1}_{\{y > x_1\}} \\ &\quad (-G_1(x_1) + G_1(y) - G_2(x_2)(x_1 - y)) \\ &\quad + (1 - \alpha)(G_1(x_1) - G_2(x_2)(x_2 - x_1) + \mathscr{G}_2(x_2)) \,, \end{aligned} \tag{2.20}$$

where G_1 is an increasing function, $\mathscr{G}_2' = G_2$ and G_2 is strictly increasing and strictly concave, are strictly consistent for $\tilde{\Theta} = (VaR_\alpha, ES_\alpha)$, $\alpha \in (0, 1)$, with respect to the class $\mathscr{P}'$. Moreover, $\mathscr{P}'$ are all real-valued probability distributions with finite mean, which have unique α-quantiles and $G_1(X)$ is integrable for all random variables X with distribution in $\mathscr{P}'$

Proof. See Corollary 5.5 in Fissler and Ziegel (2016). □

Concluding this subsection, elicitability can be seen as a useful tool for the evaluation of risk measure forecasts. The ES itself does not fulfil this statistical property, whereas the VaR does. Nevertheless, one can find strictly consistent scoring functions for the pair of ES and VaR, or more precisely the ES together with the VaR is conditional elicitable. The consecutive section is going to turn to the main topic of this thesis, the backtesting of risk measures, especially of the ES. Thus, it will also evaluate the concern if the lack of elicitability of the ES is indeed a serious drawback for the development of backtesting approaches.

3. Backtesting of Risk Measures

For the estimation of risk measure values, there are two, potentially contrary objectives both regulators and financial institutions are mainly interested in. First of all, as already outlined

above, the estimation of risk figures is closely related to the capital buffer an institution needs to put aside. On the one hand, the regulator needs to assure a sufficient magnitude of capital reserves, such that banks and insurance companies stay solvent in times of financial distress. On the other hand, financial institutions try to keep their capital buffer at the lowest possible level in order to achieve a more efficient capital allocation. Given these two adverse targets, it is essential that risk measure forecasts are sufficiently backtested against actual P&L realizations.

Correspondingly, this section is structured as follows. Subsection 3.1 introduces the general idea of backtesting. Furthermore, I want to introduce a set of rather qualitative aspects, which I believe are of particular interest for any practitioner. Moreover, these aspects are related to potential issues for both the implementation as well as the execution of risk measure backtests in real-world scenarios. Subesection 3.2 gives a short overview of the backtesting framework of VaR forecasts, whereas the subsection thereafter aims to give an overview on the discussion, whether the ES is backtestable despite its lack of elicitability.

3.1. Definition and practical aspects

Loosely speaking, as mentioned by Kratz et al. (2018), a backtest is a statistical procedure which compares forecasts of risk measures to actual realizations in order to judge whether the forcasting model is accurate or not.

First of all, in order to prevent any possible confusions, one might further categorize the type of considered forecast. In line with common practice, Emmer et al. (2015) differentiates between three different types, which are listed below.

(i) *Point forecasts* for the value of a random variable or a probability distribution. These are often expressed by an conditional expectation, i.e. $\mathbb{E}[L_t|\mathscr{F}_{t-1}]$, where again $\mathscr{F}_{t-1}$ inherits all information up to time $t-1$.

(ii) *Interval forecasts*, resulting in an interval estimate in which the forecast is expected to lie with some probability level p. Both the VaR and the ES do belong to this class of forecasts. As an example, given a loss variable L and the estimated interval (VaR_α, ∞), then the loss is expected to be within this interval with a probability of $1-\alpha$.

(iii) *Forecasts of the entire probability distribution*, like for example an estimate of the CDF or the PDF of a return loss distribution.

Therefore, both VaR and ES forecasts are in most cases categorized as interval forecasts.

For the purpose of backtesing, some notational aspects and assumptions are outlined in the following, which will be applied in the consecutive chapters.

Notation/Assumptions 3.1. *The following set of notations and assumptions in the context of backtesting is going to be used in the course of the thesis.*

- *Denote the backtesting horizon, i.e. the number of P&L observations used for the backtest, by $T \in \mathbb{N}$. The Basel Committee suggest a value of $T = 250$ days, which will be taken as a reference.*
- *Denote a vector of realized return losses over the backtesting horizon by $\vec{l} := \{l_t : t = 1,\dots,T\}$, where losses are assumed to be independently but not identically distributed. Correspondingly, the setting allows for time dependent return loss distributions.*
- *For any time $t \in \{1,\dots,T\}$ the realized loss l_t is distributed according to some unknown return loss variable L_t, i.e $l_t \sim L_t$. Furthermore, the risk manager forecasts the return loss $\widehat{L}_t$ and the related distribution $F_t(\widehat{L}_t)$ at any point in time within the backtesting horizon. It is assumed, that both L_t and $\widehat{L}_t$ follow a continuous and strictly increasing distribution for any considered time period.*
- *Denote a set of risk measure forecasts over the backtesting horizon by $\{\widehat{\rho}_t : t = 1,\dots,T\}$. For the particular case of the ES and the VaR, the estimated set of forecasts is denoted by $\{\widehat{ES}_{t,\alpha} : t = 1,\dots,T\}$ and $\{\widehat{VaR}_{t,\alpha} : t = 1,\dots,T\}$, respectively.*
- *Assume a risk manager requires a set of $k \in \mathbb{N}$ auxiliary variables at any point in time t within the backtesting horizon, either for the estimation or the backtesting of the respective risk measure. Denote the set of auxiliary variables at time t by, $(a_t^1,\dots,a_t^k) \in \mathbb{R}^k$.*

See Appendix B for a summary of the most important variables used within this thesis.

The following general definition of a backtest is motivated by Bayer and Dimitriadis (2019), nevertheless it is less strict in the sense that is allows for the inclusion of auxiliary variables in the backtest of risk measure forecasts.

Definition 3.2 (Backtest)**.** *Given the notation introduced above, a backtest of the series of forecasts $\{\hat{\rho}_t : t = 1,\dots,T\}$ for the risk measure ρ, given a time series of $k \in \mathbb{N}$ auxiliary variables, based on the realized return loss series $\vec{l}$ is a function*

$$f : \mathbb{R}^T \times \mathbb{R}^{T\times k} \times \mathbb{R}^T \to \{0,1\}\,, \tag{3.1}$$

which maps the series of forecasts, realized returns and potential auxiliary variables onto a test decisions.

In an optimal scenario no auxiliary variables are needed in order to test risk measure forecasts on P&L realizations. As an example, all common VaR backtests do not rely on any further input parameters, nevertheless this is different for ES backtests, which often require auxiliary variables like volatility or VaR estimates.

Bayer and Dimitriadis (2019) puts a lot of emphasis on developing a standalone ES backtest, which does not require any additional input parameters. Indeed, this can be seen as a criteria for the conceptual soundness of the underlying backtest, nevertheless my main focus for this thesis is on the practical applicability of backtests for financial institutions. As an example,

the estimation of both ES and VaR is tightly connected. In general, if one estimates the ES, then one can easily also obtain the corresponding VaR forecast without any further assumptions or computational difficulties. On the contrary, a backtest which requires an estimation of the entire loss distribution should be seen critical. For any non-parametric estimation method for the ES, serious additional assumptions would be needed to come up with an explicit estimate for the CDF of the underlying loss variable.
Therefore, I do generally allow for auxiliary variables in ES backtesting approaches, nevertheless I will keep a close eye on the type of auxiliary variable and if this is related with any additional assumptions or effort for the risk manager.
Up to today, there is not one single ES backtest, which can be seen as an industry standard, but, as mentioned before, there is a variety of different approaches. In addition to Definition 3.2, I believe that any backtesting approach should fulfil some further criteria in order to qualify as a standard approach for the financial industry. Therefore, I derived the following set of qualitative properties, which are relevant in real-world applications.

Proposition 3.3. *Any backtest for a series of risk measure forecasts should take into account the following qualitative criteria in order to be suitable for the implementation by both regulators and financial institutions.*

(i) Data intensity:
In the best case, a backtest should only require the actual forecasts $\hat{\rho}$ and observed realizations as input parameters. Any additional, auxiliary parameters should be easy to obtain without the need of further assumptions on the estimation model. Furthermore, regulators might have less auxiliary variables at hand as they need to rely on variables reported by financial institutions.

(ii) Computational effort:
Financial institutions in general need to apply backtests to multiple portfolios on a regular basis. Thus, a backtest which exhibits excessive computational effort might not be feasible for practical implementation. As an example, backtests which are based on a bootstrap procedure might be problematic.

(iii) Conceptual ease:
In practice, applied backtests need to be communicated with internal management, regulators and other stakeholders. Thus, the applied backtest should be conceptually sound and understandable. Furthermore, a backtest should be flexible, such that it can be aligned with changing internal or regulatory requirements.

(iv) Clear decision making:
A backtest should yield a clear decision (reject/do not reject) of a risk measure forecast at a given significance level. Optimally, one- and two-sided versions are available. Regulators care more about one-sided tests, which evaluate if the actual risk is underestimated, while from a modelling perspective two-sided tests might be more relevant.

As a consequence, all implemented backtests are not just evaluated according to classical measures like empirical size or power, but also with respect to the qualitative criteria presented in Proposition 3.3. For the remainder of this thesis, the main focus will be on one-sided risk measure tests. In practice, this is by far the more relevant case, as regulators want to assure that financial institutions do not underestimate their actual risk and thus hold to low capital buffers. Furthermore, regulators do not care about overly conservative risk forecasts by financial institutions.
The consecutive subsection is shortly going to summarize the current backtesting framework for the VaR risk measure, which was brought up by Basel Committee (n.d.-d). Indeed, the backtesting methodology for the VaR can be seen as a starting point for the development of most relevant ES backtests. Thus, the VaR backtest introduced below will also be taken up again in the chapter thereafter, where all selected ES backtesting approaches are presented.

3.2. Backtesting Value at Risk

This subsection aims to outline the current industry standard for the backtesting of the VaR, the *traffic light test*, which was introduced Basel Committee (n.d.-d). Furthermore, this subsection presents how this practical approach can be related to more theoretical contributions like Christoffersen (1998). Indeed, the current applied traffic light test is also a one-sided test, which is tailored to regulator's requirements.
Assuming a continuous loss distribution L_t, for some $t > 0$, $t \in \mathbb{N}$, it holds that

$$\mathbb{P}(L_t \geq VaR_{\alpha,t}) = 1 - \alpha \,, \tag{3.2}$$

such that the probability of a violation of the VaR is given by $1-\alpha$. As for example proposed by Christoffersen (1998), one can define the *violation indicator* of the VaR at time t as follows.

Definition 3.4 (Violation indicator)**.** *For some $t \in \{1,\dots,T\}$ and some realized return loss $l_t \sim L_t$, the violation indicator of a VaR estimate depending on the confidence level $\alpha \in (0,1)$, is defined as,*

$$I_t : (0,1) \to \{0,1\} \qquad I_t(\alpha) := \mathbb{1}_{\{l_t \geq \widehat{VaR}_{\alpha,t}\}} = \begin{cases} 0 & \text{if } l_t < \widehat{VaR}_{\alpha,t} \\ 1 & \text{if } l_t \geq \widehat{VaR}_{\alpha,t} \end{cases} . \tag{3.3}$$

Given both the identity in (3.2) and the violation indicator in Definition 3.4, Christoffersen (1998) derived the following two conditions, which need to be satisfied by any series of correct VaR forecasts.

Proposition 3.5 (VaR - Conditional Coverage)**.** *Consider a series of $T \in \mathbb{N}$ VaR forecasts and the violation indicator as defined in (3.3), then the VaR forecasts are accurate if and only if the following two conditions are fulfilled.*

(i) Unconditional coverage:
$\mathbb{E}[I_t(\alpha)] = 1-\alpha$ *for all* $t = 1,\dots,T$.

(ii) Independence condition:
$I_t(\alpha)$ *and* $I_s(\alpha)$ *are independent for all* $t \neq s$, *with* $t,s \in \{1,\dots,T\}$

If both conditions (i) and (ii) are fulfilled, then the VaR forecasts have correct conditional coverage. In this case it holds

$$I_t(\alpha) \overset{i.i.d}{\sim} Bernoulli(1-\alpha) \quad \text{for all } t \in \{1,\dots,T\}. \tag{3.4}$$

Moreover, summing up the violation indicator for all t leads to

$$\sum_{t=1}^{T} I_t(\alpha) \sim Binomial(T, 1-\alpha). \tag{3.5}$$

The Basel Committee prescribes a confidence level of $\alpha = 0.99$ as well as a backtesting period of $T = 250$, which roughly corresponds to one year if daily returns are considered. Thus, the expected number of VaR breaches is given by $\mathbb{E}\left[\sum_{t=1}^{250} I_t(0.99)\right] = 250 \cdot 0.01 = 2.5$.
For the one-sided Basel traffic light test, three colour zones are defined, which correspond to a green, a yellow and a red traffic light. Furthermore, the following set of hypothesis is evaluated in the VaR traffic light test,

$$H0: \quad \sum_{t=1}^{T} I_t(\alpha) \leq 1-\alpha \tag{3.6}$$

$$H1: \quad \sum_{t=1}^{T} I_t(\alpha) > 1-\alpha \tag{3.7}$$

Given that the forecasting model is correctly specified, the total number of violations follows a binomial distribution as outlined in (3.5). Respectively, a series of VaR estimates over a backtesting period of T observation is rejected at a significance level of κ, whenever $F^{\text{binom}}(\sum_{t=1}^{T} I_t(\alpha), T, 1-\alpha) > 1-\kappa$. One should note that $F^{\text{binom}}(X,T,p)$ denotes the cumulative probability of a value X, related to a binomial distribution with T trials and a success probability of p.
For the yellow traffic light a backtesting significance level of $\kappa = 0.05$ is applied, whereas a value of $\kappa = 0.0001$ is considered for the red traffic light. Thus, the *green zone* contains the number of violations, such that the cumulative probability of obtaining at most that many violations is below 95 %. The *yellow zone* contains the number of violations, in case the respective cumulative probability is between 95% and 99.99%. In case the cumulative probability of a certain number of violations is above 99.99%, the forecasting model is ranked in the *red zone*. If the backtest for a VaR estimation model exhibits a yellow or red traffic light, a multiplication factor is added upon the estimated VaR in order to increase the respective capital requirements and thus punish the financial institution, which understates the actual risk.[12]
Given the parameters suggested by Basel Committee (n.d.-d), i.e. $T = 250$ and $\alpha = 0.99$, the resulting traffic light zones for different numbers of violations are depicted in Table 1.

[12] See Basel Committee (n.d.-d) for more details on the procedure.

The traffic light test is a straightforward backtest, which detects financial institutions that underestimate their market risk figures and thus exhibit too many VaR violations. Nevertheless, the traffic light test has one shortcoming, as it only tests for the unconditional coverage of the VaR. Under the null hypothesis, it assumes that the occurrence of violations is independent. More precisely, it only accounts for the number of violations but not on their timing. In practice, if an estimation model is badly specified, violations might be clustered around certain events, which will not be detected by the traffic light test. As an example, Christoffersen (1998) suggests a backtest for conditional coverage of the VaR, which does also account for the independence condition stated in Proposition 3.5. Nevertheless, in practical applications the independence assumption is often checked separately in addition to the proposed traffic light test. Often, this is done by manual inspection of the occurrence of violations over time, as for example stated by Moldenhauer and Pitera (2018). Furthermore, the traffic light test does not need any auxiliary input variables in addition to VaR forecasts and realizations. The methodology is easy to understand and does not require any bootstrap procedure, such that it is also computationally efficient. Lastly, test decisions can easily be obtained at multiple significance levels of the binomial test.
Although the traffic light test exhibits theoretical shortcomings, as it is only an unconditional coverage test for the VaR, it is still the benchmark approach used within the financial industry. Moreover, the backtest satisfies all qualitative criteria outlined in Proposition 3.3. On the contrary, no comparable benchmark approach to backtest the ES, which is widely excepted amongst practitioners, has yet been agreed on.

3.3. Backtesting Expected Shortfall and the need for elicitability

The decision to introduce the ES as a primary measure of market risk was mainly based on the arguments outlined in chapter 2. Nevertheless, this inevitably calls for the need of some kind of standard backtesting procedure for the ES. Currently, the Basel Committee suggests to base capital requirements on the ES at level $\alpha = 0.975$, while still backtesting the related VaR figures at confidence levels of $\alpha = 0.99$ and $\alpha = 0.975$ by using the traffic light test outlined in the previous subsection (see Basel Committee (n.d.-b)). As for example argued by Costanzino and Curran (2018), this appears to be fairly insufficient. Furthermore, it reveals the difficulties of finding a suitable methodology, despite the multiple theoretical contributions on that topic within the last years. This subsection aims to give a short overview on the publications related to the backtesting of ES. Therefore, it also enters into the discussion, whether and how backtestability and elicitability should be related to each other.
The decision by the Basel Committee in 2013 to introduce the ES as a primary measure of market risk was often criticized, as Gneiting (n.d.), two years before, derived that the ES is not an elicitable functional. Publications like Chen (2014), thus conclude that the superior theoretical properties of the ES go hand in hand with the inability to derive suitable backtesting alternatives. Nevertheless, in the meantime contributions like Em-

Basel Traffic Light Test - VaR		
Traffic Light	**Number of Violations**	**Cumulative Probability**
Green	0	8.11%
	1	28.58%
	2	54.32%
	3	75.81%
	4	89.22%
Yellow	5	95.88%
	6	98.63%
	7	99.60%
	8	99.89%
	9	99.97%
Red	> 10	$> 99.99\%$

Table 1: Traffic light zones for the VaR backtest suggested by Basel Committee (n.d.-d) given a backtesting horizon of $T = 250$ observations and a VaR confidence level of $\alpha = 0.99$.
Source: Basel Committee (n.d.-d)

mer et al. (2015) or Acerbi and Szekely (2014) start to prevail, which argue that elicitability should not be connected with the ability to backtest a risk measure. Furthermore, also in earlier years contributions like Kerkhof and Melenberg (2004) argue that backtesting the ES is indeed not any more difficult than backtesting the VaR.

One might take the VaR as an example. As pointed out in subsection 2.4, the VaR is a elicitable risk measure, thus one can compare different forecasting models with respect to the minimization problem stated in formula (2.16). Nevertheless, a backtesting procedure does not compare multiple models, but needs to judge the accuracy of one single approach. Consequently, most VaR backtests are based on the violation indicator defined in Definition 3.4 rather than anything that has to do with elicitability. This is in line with Acerbi and Szekely (2014), which states that the property of elicitability might be useful for model selection, but not for the case of backtesting. Similar as for the VaR, also most suggested approaches to backtest the ES are not based on the concept of elicitability.[13] Therefore, the existence of a variety of potential ES backtests is another indicator that elicitability is not a necessary condition to be able to backtest a risk measure.

Furthermore, I want to add another perspective to that discussion. As for example argued in Cont et al. (2008), the risk estimation procedure might be divided into two single steps. First of all, every estimation model requires either an explicit, via a parametric approach, or an implicit, via a non-parametric approach, estimation of the return loss distribution of the underlying portfolio. Secondly, both the VaR or the ES are only a deterministic mapping from the estimated model into the real-values. If a risk manager does a proper job in modelling the underlying loss distribution, then both ES and VaR figures will be accurate. It might be more difficult to derive an untainted framework for the backtesting of the ES as it is possible for the VaR. Nevertheless, the ES is just a functional of the α-tail of return losses. Thus, any methodology to evaluate the appropriateness of the estimated tail distribution can be a reasonable backtesting tool, although it might not be a conceptually ideal backtest for the ES. My point is that just because it might be difficult to backtest the ES directly does not mean one can not judge whether a risk manager does a decent job in estimating ES forecasts. One can take the approach by Kratz et al. (2018), which will also be implemented for the purpose of this thesis, as an example. Kratz et al. (2018) approximates the ES given in Definition 2.5 by a Riemann-sum using VaR forecasts at different confidence levels between α and one. Therefore, Kratz et al. (2018) can rely on existing techniques to backtest the multiple VaR values in order to implicitly evaluate the ES forecast. This displays an indirect approach of backtesting the ES, nevertheless it might still be appropriate to evaluate ES forecasts irrespective the lack of elicitability. Overall, I belief that the lack of elicitiability of the ES should not be seen as an argument against the regulatory decision made by Basel Committee (n.d.-a).

In the following, there are three more comments I want to make on the general classifications of existing ES backtests. First of all, as already argued at the beginning of this chapter, most proposed ES backtests require additional input variables. Thus, existing approaches might be categorized according to the type of auxiliary input variables they require. As an example multiple tests rely on the VaR (Acerbi and Szekely (2014), Kratz et al. (2018)), the volatility (McNeil and Frey (2000), Nolde and Ziegel (2017)), the cumulative violation process $\int_\alpha^1 I_t(p)dp$ (Costanzino and Curran (2015), Du and Escanciano (2017)) or even the whole return loss distribution (Berkowitz (2001), Kerkhof and Melenberg (2004), Wong (2008)) as an input parameter in addition to the ES. As argued above, from a practical perspective this might be seen more or less problematic depending on the type of auxiliary variable.

[13] The ES backtests proposed by Fissler and Ziegel (2016) and Bayer and Dimitriadis (2019) can be seen as an exception as they are both based on the concept of conditional elicitability, previously introduced. Indeed, the later one will be implemented in the following chapter for the purpose of this thesis.

Secondly, there is one strand in the ES backtesting literature, which does not focus on point or interval forecast of risk measures, but rather on the backtesting of the entire estimated return loss distribution. Most of these approaches are based on *realized p-values*, which estimate the probability of observing ex-post losses based on the predicted return loss distribution. Diebold, Gunther, and Tay (1998) firstly introduced the idea of evaluating density forecasts based on realized p-values. Afterwards, both Berkowitz (2001) as well as Kerkhof and Melenberg (2004) derived ES backtests based on the concept of realized p-values. In the light of my previous argumentation, both approaches also avoid any difficulties of directly backtesting the ES. Nevertheless, both tests require an explicit estimate for the entire loss distribution, such that they are not feasible in many real-world applications. Thus, these backtests are not going to be considered in the course of this thesis.
As a third aspect, one can again differentiate between unconditional and conditional coverage tests for the ES. Given the usual parameter value of $\alpha = 0.975$, there are in general three aspects which need to be considered for a ES backtest. First of all, the number of violations exceeding the $VaR_{0.975}$ threshold, secondly the magnitude of any violation and lastly again the independence of violations beyond the $VaR_{0.975}$ value. A conditional coverage test needs to satisfy all three criteria, while an unconditional coverage test of the ES only takes into account the first two criteria. Du and Escanciano (2017) is, up to my knowledge, the only conditional coverage test for the ES that has been suggested up to now, all other approaches do not take into account the independence of VaR violations.
The following chapter is going to turn to the concrete ES backtesting approaches, which are considered within this thesis.

4. Evaluated backtests for the Expected Shortfall

This chapter presents all five backtests which are implemented in *Python* for the purpose of this thesis, as well as any adjustments which I made compared to the original approaches. As outlined in the previous chapter, there is a variety of backtests for the ES, which have been suggested within the last two decades. Although any choice of five different backtests is somehow arbitrary up to a certain extend, I tried to select five approaches with respect to the following objective.
The main objective of this thesis is to find an one-sided, unconditional coverage, backtest for the Expected Shortfall (ES) at the confidence level of $\alpha = 0.975$, which should be practically applicable with respect to the criteria stated in Proposition 3.3.

- First of all, I selected the *multinomial backtest* from Kratz et al. (2018), as it appears to be a natural extension to the binomial test for the VaR.
- Secondly, I chose the so called *"Test 2"* from Acerbi and Szekely (2014), which is probably the most prominent ES backtest and often used as a benchmark in other recent contributions.
- The third approach I selected is the *intercept ES regression (ESR) backtest* from Bayer and Dimitriadis (2019), as it takes a novel view on the ES backtesting by introducing a regression framework. Furthermore, it is the only approach presented within this thesis which is somehow related to the concept of elicitability.
- As a fourth approach the *Z-test* from Costanzino and Curran (2015) is implemented, which exploits nice distributional properties and can easily be generalized to backtest any spectral risk measure.
- The last approach I selected, is up to my knowledge the first ES backtest that has been suggested amongst researchers. Nevertheless, the *residuals bootstrap test* from McNeil and Frey (2000) is still widely used given its very intuitive concept.

4.1. Multinomial backtest from Kratz et al. (2018)

This section is first going to introduce the *multinomial backtest* brought up by Kratz et al. (2018). Whereas the original version is formulated as a two-sided test, I will propose some slight adjustments to the test decision in order to make it a 'de-facto' one-sided test, in line with the objective stated above.

4.1.1. Original approach - Kratz et al. (2018)

The idea of the multinomial backtest by Kratz et al. (2018) is relatively straightforward. As shortly mentioned in the previous chapter, one might approximate the ES by a Riemann-sum like for example in the following way,

$$ES_{t,\alpha}(L_t) \approx \frac{1}{4}[VaR_{t,\alpha} + VaR_{t,0.75\alpha+0.25} + VaR_{t,0.5\alpha+0.5} + VaR_{t,0.25\alpha+0.75}]\,. \tag{4.1}$$

Kratz et al. (2018) suggests, to implicitly backtest the ES by deriving a multinomial test for the $N = 4$ VaR figures. In order to increase the approximation accuracy, one might for example also choose $N = 8$ or $N = 16$ VaR values, i.e. quantiles of the distribution of L_t.
In the following, some notational aspects are fixed. Given $N \in \mathbb{N}$ VaR levels which are considered for the approximation of the ES in (4.1), the respective VaR confidence levels are defined as,

$$\alpha_j := \alpha + \frac{j-1}{N}(1-\alpha), \quad j = 1,\dots,N\,, \tag{4.2}$$

where α is some reference confidence level, like for example 0.975. Due to technical reasons, Kratz et al. (2018) sets $\alpha_0 := 0$ and $\alpha_{N+1} := 1$.
For any backtesting horizon $T \in \mathbb{N}$, if the respective series of VaR estimates at confidence level α_j for $j \in \{1,\dots,N\}$ has correct unconditional coverage, then according to Proposition 3.5 it holds

$$\sum_{t=1}^{T} I_t(\alpha_j) \sim Binomial(T, 1-\alpha_j). \tag{4.3}$$

Therefore, one can simultaneously test VaR estimates at all N considered confidence levels by employing a multinomial distribution. The multinomial distribution is denoted by $MN(n,(p_0,\dots,p_N))$, where each of the n trials results in

$N+1$ outcomes distributed according to the vector of success probabilities $(p_0,\dots,p_N)$.
Therefore, Kratz et al. (2018) defines the series $\{X_t\}_{t=1,\dots,T}$, by

$$X_t := \sum_{j=1}^{N} I_t(\alpha_j). \tag{4.4}$$

Moreover, X_t counts the number of breached VaR levels at time t. Furthermore, Kratz et al. (2018) uses a slight adjustment in order to consider all observations t within the backtesting horizon simultaneously. They define the so called *cell counts* by,

$$O_j = \sum_{t=1}^{T} \mathbb{1}_{\{X_t=j\}} \quad \text{for all } j=0,\dots,N\,. \tag{4.5}$$

One should note, that any violation of VaR_{t,α_j} is automatically also a violation of $VaR_{t,\alpha_{j-1}}$, as by definition $VaR_{t,\alpha_j} \geq VaR_{t,\alpha_{j-1}}$. Thus, O_j counts the number of observations over the backtesting horizon which breach the first j VaR thresholds up to the confidence level α_j, but do not breach the threshold at confidence level α_{j+1}. Given correct unconditional coverage of the VaR at all considered confidence levels α_j, the random vector $(O_0,\dots,O_N)$ is thus distributed according to the following multinomial distribution,

$$(O_0,\dots,O_N) \sim MN(T,(\alpha_1-\alpha_0,\dots,\alpha_{N+1}-\alpha_N))\,. \tag{4.6}$$

Consider the general case, where $(O_0,\dots,O_N) \sim MN(T,(\theta_1-\theta_0,\dots,\theta_{N+1}-\theta_N))$, for some arbitrary parameters $0=\theta_0<\theta_1<\ldots<\theta_N<\theta_{N+1}=1$. The formal null and alternative hypothesis according to Kratz et al. (2018) are then given by

$$\begin{aligned} H0: &\quad \theta_j = \alpha_j \quad \text{for all } j\in\{1,\dots,N\} \\ H1: &\quad \theta_j \neq \alpha_j \quad \text{for at least one } j\in\{1,\dots,N\}\,. \end{aligned} \tag{4.7}$$

Indeed, Kratz et al. (2018) evaluates three different multinomial tests for multiple amounts of VaR thresholds N. Following their conclusion, a χ^2 test based on Nass (1959) and a value of $N=8$ VaR approximation levels yields the best results and will be presented below.
Moreover, the *Nass test* test proposed by Nass (1959) is an adjustment of the standard *Pearson* χ^2 *test* introduced by Pearson (1900). As noted by Kratz et al. (2018), the Nass test is superior if cell probabilities are low, which is also the case in the considered scenario.
The test is based on the test statistic of a standard Pearson χ^2 test, depending on the choice of considered VaR levels N and the observed cell counts O_j,

$$Z_N := \sum_{j=0}^{N} \frac{(O_j - T(\alpha_{j+1}-\alpha_j))^2}{T(\alpha_{j+1}-\alpha_j)} \overset{d}{\underset{H0}{\sim}} \chi^2_N\,. \tag{4.8}$$

The Nass test consecutively uses an adjustment factor c in the following way,

$$c\cdot Z_N \overset{d}{\underset{H0}{\sim}} \chi^2_\nu, \quad \text{with } c := \frac{2\mathbb{E}[Z_N]}{Var(Z_N)} \text{ and } \nu := c\mathbb{E}[Z_N]\,, \tag{4.9}$$

where $\mathbb{E}[Z_N]=N$ and $Var(Z_N)=2N-\frac{N^2+4N+1}{T}+\frac{1}{T}\sum_{j=0}^{N}\frac{1}{\alpha_{j+1}-\alpha_j}$.
[14] Given a significance level of κ for the backtest, the null hypothesis of the two-sided test version by Kratz et al. (2018), given in (4.7), is rejected whenever $c\cdot Z_N > \chi^2_\nu(1-\kappa)$.
Given the two-sided hypothesis stated in (4.7), the multinomial ES backtest proposed by Kratz et al. (2018) does not only reject estimation models where the true risk is underestimated, but also overly conservative forecasting models. This is problematic with respect to the objective of this thesis stated above. Indeed, regulators only want to punish the underestimation of the actual risk, but not any conservative estimation approach.
For a potential example of a conservative ES estimation model, which would be rejected in the two-sided multinomial ES backtest, one might consider the following rather extreme scenario. Given a backtesting period of $T=250$ observations, the multinomial ES backtest would reject an estimation model where non of the realized return losses l_t breaches any of the estimated N VaR thresholds, i.e.

$$\begin{aligned} l_t < \widehat{VaR}_{t,\alpha_j} \qquad & \text{for all } t=1,\dots,T \\ & \text{and for all } j=1,\dots,N. \end{aligned} \tag{4.10}$$

Correspondingly, one obtains cell counts of $O_0=250$ and $O_j=0$ for all $j=1,\dots,N$, which leads to a rejection in the two-sided multinomial backtest.
Nevertheless, by design the two-sided version proposed by Kratz et al. (2018) only has few room to reject conservative ES estimation models. Thus, on the conservative side, the multinomial ES backtest only detects cases where a risk manager extremely overestimated the underlying risk. Especially, for rather short backtesting periods like $T=250$, the multinomial backtest will only reject extreme examples, like the one depicted above. Nevertheless, I want to add a slight adjustment to the test decision in order to make it indeed a de-facto one-sided test for any reasonable number of backtesting observations T.

4.1.2. Adjusted approach - De-facto one-sided test

In order to be able to compare different backtests with respect to the objective stated above, I want to propose a de-facto one sided test, with the following null and alternative hypothesis,

$$\begin{aligned} H0: &\quad \theta_j \geq \alpha_j \quad \text{for all } j\in\{1,\dots,N\} \\ H1: &\quad \theta_j < \alpha_j \quad \text{for at least one } j\in\{1,\dots,N\}\,. \end{aligned} \tag{4.11}$$

For the de-facto one sided test I add an additional criteria in the form of a *conservatism indicator* to the test decision. The conservatism indicator is defined as

$$\mathbb{1}_{\text{conservatism}} := \begin{cases} 0 & \text{if } \exists j\in\{1,\dots,N\}: \quad \sum_{t=1}^{T} I_t(\alpha_j) > 1-\alpha_j \\ 1 & \text{if } \forall j\in\{1,\dots,N\}: \quad \sum_{t=1}^{T} I_t(\alpha_j) \leq 1-\alpha_j \end{cases}, \tag{4.12}$$

where $I_t(\alpha)$ is again the violation indicator as defined in (3.2). Moreover, the conservatism indicator is equal to one, if for all considered N VaR levels the number of observed violations is

[14] See Nass (1959) for more details on the methodology of the proposed test.

smaller or equal than the expected number of violations, given the model is correctly specified. In this case the ES estimation model is definitely on the conservative side and should therefore not be rejected in a one-sided test. Accordingly, the null hypothesis in (4.11) is rejected if $c \cdot Z_N > \chi^2_\nu(1-\kappa)$ and $\mathbb{1}_{\text{conservatism}} = 0$.

As previously outlined, only extremely conservative ES estimation models are rejected by the two-sided approach and this only in case a rather long backtesting period is considered. The proposed adjustment of the test decision additionally also rules out the rejection of those, obviously too conservative, models. Therefore, it turns the two-sided test proposed by Kratz et al. (2018) into a de-facto one-sided version at least for any reasonable backtesting horizon and all relevant significance levels of the backtest, like $\kappa = 0.05$ or $\kappa = 0.0001$.

Definitely, the proposed version might not be a conceptual ideal backtest, in the sense that the stipulated significance level κ might not perfectly coincide with the actual significance level of the one-sided test. Nevertheless, given the argumentation above, the impact of the proposed adjustment is expected to be rather marginal. More importantly, the adjusted multinomial backtest can also be compared to the other one-sided ES backtests, which are presented in the following subsections. With respect to the criteria stated in Proposition 3.2, the backtest requires the vector of VaR estimates, $(\widehat{VaR}_{t,\alpha_1}, \ldots, \widehat{VaR}_{t,\alpha_N})$, and P&L realizations as input variables. Although it does not backtest the ES directly, the related VaR figures in general can be easily obtained. Furthermore, given the multinomial distribution of the test statistic, no bootstrap procedure is required and the methodology is an intuitive generalization of the binomial test used for the VaR. Although, there is no straightforward one-sided version of the original test methodology proposed by Kratz et al. (2018), the backtest might still be highly relevant for practical applications given its easy concept and the requirement of both few input parameters and few computational effort.

4.2. "Test 2" from Acerbi and Szekely (2014)

The second test evaluated within this thesis stems from Acerbi and Szekely (2014) and is often referred to as *"Test 2"* within related contributions. Indeed, it is the second out of three backtests proposed by Acerbi and Szekely (2014). This subsection aims to introduce the test methodology and shortly discuss the need of a bootstrap algorithm.

The backtest suggested by Acerbi and Szekely (2014) is based on the following unconditional expectation

$$ES_{t,\alpha}(L_t) = \mathbb{E}\left[\frac{L_t I_t(\alpha)}{1-\alpha}\right], \tag{4.13}$$

which is a correct specification of the ES, as the loss distribution L_t is assumed to be continuous. Depending on the vector of observed losses, $\vec{l}$, Acerbi and Szekely (2014)[15] defines the test statistic,

$$Z := \left(\sum_{t=1}^{T} \frac{l_t I_t(\alpha)}{T(1-\alpha)\widehat{ES}_{t,\alpha}}\right) - 1\,. \tag{4.14}$$

Indeed, Acerbi and Szekely (2014) proposes a one-sided test, which ought to detect whether the estimated risk, $\widehat{ES}_{t,\alpha}$, understates the actual risk given by $ES_{t,\alpha}$. With respect to the test statistic Z defined in (4.14), the following expectations can be derived, under the assumptions that the ES estimation model is firstly correctly specified or secondly underestimates the actual risk.

Proposition 4.1 (Acerbi and Szekely (2014), Proposition A.3)**.** *Given the test statistic Z defined in equation (4.14) it holds,*

(i) $\mathbb{E}[Z] = 0$*, given that the ES estimation model is correctly specified, and*

(ii) $\mathbb{E}[Z] > 0$*, given that the ES estimation model underestimates the actual underlying risk.*

Proof. (i) Under the assumption that the estimate $\widehat{ES}_{t,\alpha}$ is correctly specified for all t within the backtesting horizon, the identity (4.13) yields,

$$\begin{aligned}
\widehat{ES}_{t,\alpha} &= \mathbb{E}\left[\frac{l_t I_t(\alpha)}{1-\alpha}\right] \\
&\Leftrightarrow \quad \mathbb{E}\left[\frac{l_t I_t(\alpha)}{1-\alpha}\frac{1}{\widehat{ES}_{t,\alpha}}\right] = 1 \\
&\Leftrightarrow \quad \mathbb{E}\left[\frac{l_t I_t(\alpha)}{1-\alpha}\frac{1}{\widehat{ES}_{t,\alpha}}\right] - 1 = 0\,,
\end{aligned} \tag{4.15}$$

for all $t \in \{1, \ldots, T\}$. Furthermore, for the test statistic Z it holds,

$$\begin{aligned}
\mathbb{E}[Z] &= \mathbb{E}\left[\left(\sum_{t=1}^{T} \frac{l_t I_t(\alpha)}{T(1-\alpha)\widehat{ES}_{t,\alpha}}\right) - 1\right] \\
&= \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\frac{l_t I_t(\alpha)}{1-\alpha}\frac{1}{\widehat{ES}_{t,\alpha}}\right] - 1 \\
&\overset{(4.15)}{=} 0\,.
\end{aligned} \tag{4.16}$$

(ii) Given that the estimated ES model underestimates the true risk, the actual ES values $ES_{t,\alpha}$ are larger or equal to the estimated risk figures $\widehat{ES}_{t,\alpha}$. Furthermore, there exists some t within the backtesting horizon, such that $ES_{t,\alpha} > \widehat{ES}_{t,\alpha}$. Therefore under the assumption of risk

[15] Note that the notation is slightly different in this thesis compared to Acerbi and Szekely (2014) given the different sign conventions used for the definition of the ES.

underestimation it holds,

$$\begin{aligned}\mathbb{E}[Z] &= \mathbb{E}\left[\left(\sum_{t=1}^{T}\frac{l_t I_t(\alpha)}{T(1-\alpha)\widehat{ES}_{t,\alpha}}\right)-1\right] \\ &= \frac{1}{T}\sum_{t=1}^{T}\underbrace{\mathbb{E}\left[\frac{l_t I_t(\alpha)}{1-\alpha}\right]}_{=ES_{t,\alpha}}\frac{1}{\widehat{ES}_{t,\alpha}}-1 \\ &= \frac{1}{T}\sum_{t=0}^{T}\underbrace{\frac{ES_{t,\alpha}}{\widehat{ES}_{t,\alpha}}}_{\geq 1}-1 \\ &> 0\,, \end{aligned} \tag{4.17}$$

whereas the last inequality holds under the assumption that the ES model underestimates the actual risk.

□

Given the expectations derived in Proposition 4.1, one can derive the following null and alternative hypothesis for the Test 2 inspired by Acerbi and Szekely (2014)

$$\begin{aligned} H0: &\quad Z \leq 0 \\ H1: &\quad Z > 0\,. \end{aligned} \tag{4.18}$$

Indeed, the true distribution of the test statistic under the null hypothesis is unknown, thus one can not exploit any distributional properties to decide whether a realization of the test statistic is indeed significantly larger than zero. Acerbi and Szekely (2014) suggests the use of a simulation to determine the the p-value of the backtest related to some vector of realizations $\vec{l}$. Therefore, one needs to store all estimated return loss variables $\widehat{L}_t$ in order to simulate the distribution of Z under the null hypothesis in the following way.

- Calculate the value of the test statistic related to the observed losses, i.e. $Z(\vec{l})$.
- Simulate M independent bootstrap trials for every point in time within the backtesting horizon from the estimated loss distributions, i.e simulate $l_t^i \sim \widehat{L}_t$ for $i = 1,\ldots,M$ and $t = 1,\ldots,T$.
- Calculate the value of the test statistic related to any bootstrap trial, i.e. $Z^i = Z(\vec{l^i})$, where $\vec{l^i} = \{l_1^i,\ldots,l_T^i\}$ for $i = 1,\ldots,M$.
- Estimate the p-value related to the vector of observed losses $\vec{l}$ given by,

$$p = \frac{1}{M}\sum_{j=1}^{M}\mathbb{1}_{\{Z^i > Z(\vec{l})\}}\,. \tag{4.19}$$

Given the estimated p-value and a significance level κ for the backtest, the null hypothesis is rejected whenever $p < \kappa$.
The simulation procedure suggested by Acerbi and Szekely (2014) is necessary as the distribution of the test statistic under the null hypothesis is unknown. Moreover, the distribution in general also depends on the fitted model for the underlying portfolio, i.e. on the estimation of $\widehat{L}_t$ and $\widehat{ES}_{t,\alpha}$, respectively. This might be seen as a drawback of the outlined backtest, with respect to the criteria stated in Proposition 3.3, for two particular reasons. First of all, the bootstrap procedure is related to an increased computational effort. Secondly, the simulation requires the risk manager to store the estimated return loss $\widehat{L}_t$ for all observations within the backtesting horizon.
One reason why the test outlined above is still widely applied, is that Acerbi and Szekely (2014) recognizes that critical values for the test decision are indeed stable for different applied estimation models. Acerbi and Szekely (2014) analyses the distribution of the test statistic given that the estimated loss $\widehat{L}_t$ follows a standard normal distribution as well as multiple t-distributions with varying degrees of freedom. Given the backtest significance levels of $\kappa = 0.05$ and $\kappa = 0.0001$, stipulated by the Basel Committee, Acerbi and Szekely (2014) reports the following critical values based on their simulation study, as outlined in Table 2.
Indeed, critical values are comparable for estimated loss distributions which follow a standard normal or a t-distributions with at least $\nu = 5$ degrees of freedom. Acerbi and Szekely (2014) argues that a t-distribution with $\nu = 3$ degrees of freedom corresponds to a extremely heavy tailed loss distribution which is rather uncommon for actual portfolio losses. Furthermore, they argue that the backtest would be more penalizing in that case and thus still reject any model which underestimates the actual risk. Overall, they conclude that critical values for the backtesting significance levels of $\kappa = 0.05$ and $\kappa = 0.0001$ are fairly stable at values of 0.7 and 1.8, respectively. Following their argumentation, it is therefore not necessary to save estimated return losses $\widehat{L}_t$ and apply a bootstrap, as realized test statistic values can immediately be compared to the proposed critical values.
In order to complement the analysis by Acerbi and Szekely (2014) on the robustness of critical values, I fitted two different ES estimation models on log-return losses of the S&P 500 and simulated the respective distributions of the test statistic under the null hypothesis. The simulation is based on the bootstrap procedure outlined above and $M = 10000$ simulation trials. The respective analysis is deferred to appendix A attached to this thesis.
Based on the simulated critical values depicted in Table 2 as well as on my analysis results outlined in appendix A, I believe that critical values are reasonably stable at a backtesting confidence level of $\kappa = 0.05$. Nevertheless, for more extreme quantiles of the test statistic distribution, like example $\kappa = 0.0001$, the usage of fixed critical values might lead to inaccurate test decisions.
Concluding, depending on the underlying portfolio and the type of estimation model a risk manager who wants to apply the Test 2 from Acerbi and Szekely (2014) faces the following trade-off. On the one hand, using fixed critical values leads to a superior computational performance and fewer input variables on the other hand the abandonment of a bootstrap procedure might lead to imprecise decisions in certain situations. Taking into

	Critical values for different backtest significance levels	
Estimated return loss distribution	5 %	0.01 %
t_3	0.82	4.4
t_5	0.74	2.0
t_{10}	0.71	1.9
t_{100}	0.70	1.8
standard Normal	0.70	1.8

Table 2: Simulation based critical values of the test statistic Z at both backtest significance levels required by the Basel Committee, for estimation models based on the standard normal distribution and multiple standard t-distributions with varying degrees of freedom, t_ν.

Source: Acerbi and Szekely (2014)

account this potential trade-off, I decide to apply Test 2 proposed by Acerbi and Szekely (2014) with a fixed critical value of $Z^* = 0.70$ given a backtesting significance level of $\kappa = 0.05$. In scenarios with a backtesting significance level of $\kappa = 0.0001$, which corresponds to the red zone in the VaR traffic light test, I decide to rely on a bootstrap procedure, as critical values show a high divergence for this rather extreme quantile of simulated test statistic distributions. Moreover, any bootstrap test decision of the Test 2 will be based on $M = 1000$ simulation trials in the consecutive chapters.

4.3. Intercept ES regression backtest from Bayer and Dimitriadis (2019)

This subsection presents an adjusted version of the ES backtest derived in Bayer and Dimitriadis (2019). Their approach is based on a regression framework, which takes a novel view on the backtesting issue of the ES. Indeed, the approach is based on the results regarding the conditional elicitability of ES, derived by Fissler and Ziegel (2016) and outlined in Proposition 2.16 within this thesis. Similar to the idea of fitting a quantile regression for the VaR, Bayer and Dimitriadis (2019) proposes a framework to estimate a regression for the ES of return losses distributed by L_t on estimated ES figures $\widehat{ES}_{t,\alpha}$ as explanatory variables. Moreover, taking into account the different sign conventions within this thesis, they propose to estimate the following regression,

$$-ES_{t,\alpha} = -ES_\alpha(L_t|\mathscr{F}_{t-1}) = \gamma_0 - \gamma_1 \widehat{ES}_{t,\alpha} \,. \tag{4.20}$$

Bayer and Dimitriadis (2019) outlines two different test specifications a two-sided version, which is labelled as the *ESR (ES regression) backtetst* and a one-sided *intercept ESR backtest*. In the following, an adjusted version of the one-sided intercept ESR backtest will be presented. Therefore, Bayer and Dimitriadis (2019) sets the slope parameter in (4.20) equal to one and obtains

$$-ES_{t,\alpha} = -ES_\alpha(L_t|\mathscr{F}_{t-1}) = \gamma_0 - \widehat{ES}_{t,\alpha} \,. \tag{4.21}$$

Given the ES forecasting model is perfectly accurate, i.e. $ES_{t,\alpha} = \widehat{ES}_{t,\alpha}$ the estimated intercept parameter γ_0 will equal to 0. In case of an overly conservative forecasting model, one will obtain $\gamma_0 > 0$. On the contrary, an estimation model which understates the actual risk leads to a situation where $\widehat{ES}_{t,\alpha} < ES_{t,\alpha}$ for a reasonable number of observations within the backtesting period, such that an estimated intercept of $\gamma_0 < 0$ is to be expected. Therefore, formally the following one-sided null and alternative hypotheses are going to be evaluated

$$\begin{aligned} H0: &\quad \gamma_0 \geq 0 \\ H1: &\quad \gamma_0 < 0 \,. \end{aligned} \tag{4.22}$$

4.3.1. Regression estimation based on conditional elicitability

As outlined by Bayer and Dimitriadis (2019), one of the main difficulties is to consistently estimate a regression for the ES of a series of return losses $l_t \sim L_t$ on ES forecasts $\widehat{ES}_{t,\alpha}$. As mentioned in Proposition 2.15 within this thesis, the ES itself is not an elicitable functional, thus there exists no strictly consistent scoring function in the sense of Definition 2.13. Therefore, there is no potential objective function for any *Maximum-Likelihood (ML) estimation* procedure, in order to directly estimate γ_0 in regression (4.21).

The solution proposed by Bayer and Dimitriadis (2019) is to exploit the conditional elicitability of the vector $(VaR_{t,\alpha}, ES_{t,\alpha})$ in order to simultaniously estimate two regression equations for the quantile and for the ES of the return losses $l_t \sim L_t$. Formally, the regression system is given by,

$$\begin{aligned} -l_t &= \beta_0 - \widehat{VaR}_{t,\alpha} + \varepsilon_t^q \,, \\ -l_t &= \gamma_0 - \widehat{ES}_{t,\alpha} + \varepsilon_t^e \,, \end{aligned} \tag{4.23}$$

where $q_\alpha(\varepsilon_t^q|\mathscr{F}_{t-1}) = 0$ and $ES_\alpha(\varepsilon_t^e|\mathscr{F}_{t-1}) = 0$ almost surely. As conditional VaR and ES forecasts are considered at any time t based on the σ-Algebra $\mathscr{F}_{t-1}$, the conditions proposed to the error terms are equivalent to,

$$\begin{aligned} -VaR_\alpha(L_t|\mathscr{F}_{t-1}) &= \beta_0 - \widehat{VaR}_{t,\alpha} \,, \\ -ES_\alpha(L_t|\mathscr{F}_{t-1}) &= \gamma_0 - \widehat{ES}_{t,\alpha} \,. \end{aligned} \tag{4.24}$$

Consequently, one can use realized return losses l_t as well as $\widehat{VaR}_{t,\alpha}$ and $\widehat{ES}_{t,\alpha}$ estimates over the backtesting horizon in order to fit the regression system (4.23) by making use of a suitable strictly consistent scoring function as outlined in Proposition 2.16.

Indeed, Bayer and Dimitriadis (2019) puts a major focus on developing a standalone backtest for the ES. Therefore, they replace $\widehat{VaR}_{t,\alpha}$ in the regression system (4.23) by forecasts of the ES, $\widehat{ES}_{t,\alpha}$. As a consequence, they only require ES forecasts and realized losses as an input, but end up with a certain degree of model misspecification.[16] Nevertheless, the main objective of this thesis is to develop a practically applicable backtest. As previously argued, adding VaR forecasts as additional input parameters for the backtest is not considered as a major drawback in this regard. Thus, in order to simplify the methodology of the applied backtest I propose to estimate the regression system as outlined in (4.23) and therefore deviate from the approach suggested by Bayer and Dimitriadis (2019). For the regression estimation, the following strictly consistent scoring function for the pair of $(VaR_{t,\alpha}, ES_{t,\alpha})$ is considered in line with Bayer and Dimitriadis (2019).[17]

Proposition 4.2. *Let t be any observation within the backtesting horizon, i.e. $t \in \{1,\dots,T\}$. Denote the VaR and ES forecast at time t by $\widehat{VaR}_{t,\alpha}$ and $\widehat{ES}_{t,\alpha}$, respectively. Furthermore, denote the realized return loss by l_t and define the vector of regression parameters by $\theta := (\beta_0, \gamma_0)$. Then the scoring function S at time t defined through,*

$$\begin{aligned} &S(\widehat{VaR}_{t,\alpha}, \widehat{ES}_{t,\alpha}, l_t, \theta) := \\ &\frac{1}{\gamma_0 - \widehat{ES}_{t,\alpha}} \cdot \Bigg((\gamma_0 - \widehat{ES}_{t,\alpha}) - (\beta_0 - \widehat{VaR}_{t,\alpha}) \\ &+ \frac{\mathbb{1}_{\{l_t > \beta_0 - \widehat{VaR}_{t,\alpha}\}} \cdot (\beta_0 - \widehat{VaR}_{t,\alpha} - l_t)}{1-\alpha} \Bigg) \\ &- \log(-(\gamma_0 - \widehat{ES}_{t,\alpha})) \,, \end{aligned} \tag{4.25}$$

is strictly consistent for the pair of $(VaR_{t,\alpha}, ES_{t,\alpha})$ in the sense of Definition 2.13.

Proof. Using the notation of Proposition 2.16, define

$$\begin{aligned} x_{1,t} &:= (\gamma_0 - \widehat{ES}_{t,\alpha}) \,, \\ x_{2,t} &:= (\beta_0 - \widehat{VaR}_{t,\alpha}) \,, \\ y_t &:= l_t \,, \\ G_1(x) &:= 0 \,, \\ \mathscr{G}_2(x) &:= -\log(-x) \,, \end{aligned}$$

for t within the backtesting horizon. Furthermore, it holds $G_2(x) = \mathscr{G}_2'(x) = -1/x$, such that G_1 is an increasing function and G_2 is both strictly concave and strictly increasing. Moreover, according to the result of Proposition 2.16, the following scoring function is strictly consistent for the pair of $(VaR_{t,\alpha}, ES_{t,\alpha})$.

$$\begin{aligned} &S(x_{1,t}, x_{2,t}, y_t) = \\ &\mathbb{1}_{\{y_t > x_{1,t}\}}(-G_1(x_{1,t}) + G_1(y_t) - G_2(x_{2,t})(x_{1,t} - y_t)) \\ &+(1-\alpha)(G_1(x_{1,t}) - G_2(x_{2,t})(x_{2,t} - x_{1,t})) + \mathscr{G}_2(x_{2,t}) \\ &= \mathbb{1}_{\{y_t > x_{1,t}\}} \left(-\frac{1}{-x_{2,t}} \cdot (x_{1,t} - y_t) \right) \\ &+(1-\alpha) \left(-\frac{1}{-x_{2,t}} (x_{2,t} - x_{1,t}) \right) - \log(-(x_{2,t})) \\ &= \frac{1}{x_{2,t}} \left((x_{2,t} - x_{1,t}) + \frac{\mathbb{1}_{\{y_t > x_{1,t}\}}(x_{1,t} - y_t)}{1-\alpha} \right) \\ &- \log(-x_{2,t}) \end{aligned} \tag{4.26}$$

Inserting $x_{1,t}$, $x_{2,t}$ and y_t into (4.26) yields the desired scoring function outlined in (4.25). □

Following the estimation procedure described by Bayer and Dimitriadis (2019), the ML-estimator of the regression parameter $\theta = (\beta_0, \gamma_0)$ related to the regression system (4.23) is given by,

$$\hat{\theta}_T = \arg\min_{\theta \in \Theta} \frac{1}{T} \sum_{t=1}^{T} -S(\widehat{VaR}_{t,\alpha}, \widehat{ES}_{t,\alpha}, l_t, \theta) \,, \tag{4.27}$$

where the strictly consistent scoring function S defined in formula (4.25) is used.[18] Following the regression estimation, one can test the resulting parameter value γ_0 with respect to the hypotheses stated in (4.22).

4.3.2. Bootstrap test decision

Bayer and Dimitriadis (2019) suggests to apply a Wald-type test statistic based on some consistent covariance estimator $\hat{\Omega}$. Both Bayer and Dimitriadis (2019) and Dimitriadis and Bayer (2019) can be taken as a reference for the methodology and the implementation of a consistent asymptotic covariance estimator $\hat{\Omega}$, which is applied for their proposed test decision. Nevertheless, both the theory behind the estimator itself as well as the implementation procedure is highly advanced and might thus be burdensome for the practical implementation by financial institutions.

Therefore, I propose to rely on a bootstrap procedure. Although this does increase the computational effort, it greatly simplifies the methodology behind the applied backtest.

The idea is to estimate γ_0 related to a set of T VaR and ES forecasts and a set of actual return loss realizations $\vec{l}$. In order to obtain the distribution of γ_0 under the assumption that the model is correctly specified, M bootstrap simulation trials are applied. The overall procedure is as follows.

- Estimate the values of γ_0 and β_0 related to the observed losses and the given risk estimates, i.e. estimate $\gamma_0(\vec{l})$ and $\beta_0(\vec{l})$.

[16] See Bayer and Dimitriadis (2019) chapter 2.4 for asymptotic theory on the estimation procedure under model misspecification.

[17] Again, note that the scoring function slightly differs compared to Bayer and Dimitriadis (2019) given the divergent sign conventions.

[18] Dimitriadis and Bayer (2019) derives both consistency and asymptotic normality of the ML-estimator outlined above in case the parametric model given in (4.23) is correctly specified.

- In order to be able to simulate under the assumption of a correct estimation model, adjust the risk estimates by $\widehat{VaR}^{\text{adj}}_{t,\alpha} := \widehat{VaR}_{t,\alpha} - \beta_0(\vec{l})$ and $\widehat{ES}^{\text{adj}}_{t,\alpha} := \widehat{ES}_{t,\alpha} - \gamma_0(\vec{l})$.
- For any bootstrap trial $i \in \{1,\dots,M\}$, draw T triples $(\widehat{VaR}^{\text{adj}}_{t,\alpha}, \widehat{ES}^{\text{adj}}_{t,\alpha}, l_t)$ with replacement from the actual sample. Estimate the regression parameter γ_0^i based on the respective bootstrap sample for $i = 1,\dots,M$.
- Estimate the p-value related to the vector of observed losses $\vec{l}$ given by,

$$p = \frac{1}{M}\sum_{i=1}^{M} \mathbb{1}_{\{\gamma_0^i < \gamma_0(\vec{l})\}} \,. \tag{4.28}$$

Again, given some significance level for the backtest κ, the null hypothesis stated in (4.22) is rejected whenever $p < \kappa$. Moreover, in the following chapters, I will conduct the intercept ESR backtest based on $M = 100$ bootstrap simulations.[19]
Concluding, the intercept ESR backtest proposed within this thesis differs from the original version derived by Bayer and Dimitriadis (2019) with respect to the following two aspects. First of all it uses both VaR and ES forecasts as input variables and thus avoids model misspecification. Secondly, the test decision is based on a bootstrap procedure. Therefore, no cumbersome derivation and no further assumptions are required in order to come up with a consistent asymptotic covariance estimator. With respect to the criteria stated in Proposition 3.3, especially the bootstrap decision might be seen as a drawback of the outlined backtest, as it is related to high computation times. Furthermore, the overall test methodology is rather complex and requires a decent understanding of the concept of conditional elicitability. Nevertheless, the testing framework is highly interesting, as it provides a new and fundamentally different view on the issue of backtesting the ES compared to all previously introduced approaches.

4.4. Z-Test from Costanzino and Curran (2015)

This section is going to present the ES backtest proposed by Costanzino and Curran (2015). Moreover, the approach can be applied for any risk measure which belongs to the class of spectral risk measures according to definition Definition 2.7. The following subsection is first going to introduce the methodology of the backtest outlined by Costanzino and Curran (2015) and Costanzino and Curran (2018), whereas the sub-section thereafter is going to present a slightly adjusted, approximative version which requires less input variables.

[19] Note that a value of only $M = 100$ simulation trials is applied, as a ML estimation needs to be conducted in every simulation trial, which leads to a fast increase of computational time depending on the choice of M. Nevertheless, this choice of M still leads to satisfying results as outlined in the following chapters.

4.4.1. Original approach - Costanzino and Curran (2015)

As an extension of the violation indicator $I_t(\alpha)$ defined for the VaR, Costanzino and Curran (2015) suggests to define the so called *spectral risk measure violation rate* as follows.

Definition 4.3 (Spectral risk measure violation rate)**.** *Let ψ be an admissible risk spectrum in the sense of Definition 2.6, then for any t within the backtesting horizon, $X^t_\psi \in [0,1]$ is defined as,*

$$X^t_\psi := \int_0^1 \psi(p) I_t(p) dp \,. \tag{4.29}$$

Moreover, for any backtesting horizon T, define the spectral risk measure violation rate, $X^T_\psi \in [0,1]$ for an admissible risk spectrum ψ as,

$$X^T_\psi := \frac{1}{T}\sum_{t=1}^{T} X^t_\psi = \frac{1}{T}\sum_{t=1}^{T}\int_0^1 \psi(p) I_t(p) dp \,. \tag{4.30}$$

Compared to the violation indicator $I_t(\alpha)$ for the VaR, the spectral risk measure violation rate measures the exceedance of any VaR level, where the respective confidence level lies within the support of ψ. Therefore, in case of the ES, the spectral risk measure violation rate takes into account both the amount and the magnitude of losses beyond the VaR threshold with confidence level α. Whereas $I_t(\alpha)$ can only take on the binary values 0 or 1, the spectral risk measure violation rate takes on continuous values in the interval $[0,1]$. Given that the underlying spectral risk measure is correctly specified, $\{X^t_\psi\}_{t=1}^T$ are i.i.d. distributed and furthermore, $\mathbb{P}[L_t \geq VaR_{t,p}] = 1-p$ for all $p \in \text{supp}\ \psi$ and $t \in \{1,\dots,T\}$. Given the model is accurately estimated, Costanzino and Curran (2015) derives the following distributional properties of the spectral risk measure violation rate.

Proposition 4.4 (Costanzino and Curran (2015))**.** *Given a correctly specified spectral risk measure based on some admissible risk spectrum ψ, then the mean and variance of the spectral risk measure violation rate X^T_ψ are given by,*

$$\mu_\psi := \mathbb{E}[X^T_\psi] = \int_0^1 \psi(p)(1-p)dp \,, \tag{4.31}$$

$$\sigma^2_\psi := Var[X^T_\psi] = \frac{1}{T}\left(2\int_0^1\int_0^p \psi(p)\psi(q)(1-q)dqdp - \mu^2_\psi\right) \,. \tag{4.32}$$

Moreover, under the assumption that the spectral risk measure is correctly specified, the spectral risk measure violation rate is asymptotically normally distributed with,

$$Z^T_\psi := \frac{X^T_\psi - \mu_\psi}{\sigma_\psi} \xrightarrow[T\to\infty]{d} N(0,1) \,. \tag{4.33}$$

Proof. (i) The derivation of both moments of the spectral risk measure violation rate μ_ψ and σ_ψ follows according to Proposition 3.4 in Costanzino and Curran (2015) taking into account the different conventions within this thesis. Note for example, that $\mathbb{P}[L_t \geq VaR_{t,p}] = 1-p$ given the notation used within this thesis, which explains the difference in (4.31) and (4.32) compared to Costanzino and Curran (2015).

(ii) As derived in Lemma 3.5 within Costanzino and Curran (2015), the asymptotic distribution of X_{ψ}^{T} follows by the *Lindeberg-Levy central limit theorem* as both μ_{ψ} given in (4.31) and σ_{ψ} given in (4.32) are bounded.

□

For the particular case of backtesting the ES, one can construct a one-sided Z-test by exploiting the distributional properties of $Z_{\psi_{ES}}^{T}$. As pointed out in Lemma 2.8 within this thesis, the admissible risk spectrum related to the ES is given by $\psi_{ES}(p) = \frac{1}{1-a}\mathbb{1}_{\{\alpha \leq p \leq 1\}}$. Applying ψ_{ES} to the previous derivations leads to the consecutively listed results, following both Costanzino and Curran (2015) and Costanzino and Curran (2018).[20]

- According to Costanzino and Curran (2018), the spectral risk measure violation rate for the ES can be calculated by,

$$\begin{aligned} X_{ES_\alpha}^{t} &= \int_0^1 \psi_{ES}(p) I_t(p) dp = \frac{1}{1-\alpha}\int_\alpha^1 I_t(p)dp \\ &= \left(1 - \frac{F_t(-L_t)}{1-\alpha}\right) I_t(\alpha)\,, \qquad (4.34) \\ X_{ES_\alpha}^{T} &= \frac{1}{T}\sum_{t=1}^{T} X_{ES_\alpha}^{t} = \frac{1}{T\cdot(1-\alpha)}\sum_{t=1}^{T}\int_\alpha^1 I_t(p)dp \\ &= \frac{1}{T}\sum_{t=1}^{T}\left(1 - \frac{F_t(-L_t)}{1-\alpha}\right) I_t(\alpha)\,. \qquad (4.35) \end{aligned}$$

  Note that in equation (4.34) and (4.35), the term $1 - F_t(-L_t)/(1-\alpha)$ accounts for the severity of a VaR breach beyond the confidence level of α.

- Given that the ES is correctly specified, the following moments of the ES violation rate are derived in Costanzino and Curran (2015),

$$\begin{aligned} \mu_{ES_\alpha} &:= \mathbb{E}[X_{ES_\alpha}^{T}] = \frac{1-\alpha}{2}\,, \qquad (4.36) \\ \sigma_{ES_\alpha}^{2} &:= Var[X_{ES_\alpha}^{T}] = \\ &\quad \frac{1-\alpha}{T}\left(\frac{4-3(1-\alpha)}{12}\right)\,. \qquad (4.37) \end{aligned}$$

- Given the two previous listings and Proposition 4.4, the following distribution of the ES violation rate can be defined,

$$\begin{aligned} \lim_{T\to\infty} Z_{ES_\alpha}^{T} &:= \frac{X_{ES_\alpha}^{T} - \mu_{ES_\alpha}}{\sigma_{ES_\alpha}} \qquad (4.38) \\ &= \sqrt{3T}\left(\frac{2X_{ES_\alpha}^{T} - \alpha}{\sqrt{\alpha(4-3\alpha)}}\right) \sim N(0,1)\,. \end{aligned}$$

[20] Again note that some minor adjustments need to be made compared to Costanzino and Curran (2015) and Costanzino and Curran (2018) given slightly different conventions within this thesis.

Based on (4.35) and (4.38), one can calculate the realized value of both $X_{ES_\alpha}^{T}(\vec{l})$ and $Z_{ES_\alpha^T}(\vec{l})$, related to some set of return loss observations $\vec{l}$. Obviously, a high value of $Z_{ES_\alpha}^{T}(\vec{l})$ also indicates a high value of the ES violation rate and thus a potential underestimation of the actual risk. In order to test the one-sided hypothesis that the modelled risk $\widehat{ES}_{t,\alpha}$ underestimates the actual risk given by $ES_{t,\alpha}$, Costanzino and Curran (2018) suggests a similar procedure as for the VaR traffic light. More precisely, Costanzino and Curran (2018) argues that the realized ES violation rate, i.e. $X_{ES_\alpha}^{T}(\vec{l})$, can be seen as the equivalent to the number of observed violations in the VaR traffic light test. The higher the realized ES violation rate, the higher is the degree of risk underestimation.[21]

Denote the realized ES violation rate by $X_{ES_\alpha}^{T}(\vec{l}) = x$. Then the underlying estimation model is rejected in the one-sided test, at a significance level of κ, if the cumulative probability of $X_{ES_\alpha}^{T}$ being lower or equal to x is larger than $1-\kappa$. This is indeed a similar test decision as outlined for the VaR traffic light test in chapter 3.2 within this thesis. Formally the estimation model is rejected whenever,

$$\mathbb{P}[X_{ES_\alpha}^{T} \leq x] > 1 - \kappa\,. \qquad (4.39)$$

In order to achieve a concrete test decision, on can standardize $X_{ES_\alpha}^{T}$ by μ_{ES_α} and σ_{ES_α} and exploit the distributional properties outlined in (4.38). Applying the approximate distribution of $Z_{ES_\alpha}^{T}$ under the assumption that the model is correctly specified, the estimation model in the one-sided test, related to some set of realized losses $\vec{l}$, is rejected whenever,

$$Z_{ES_\alpha}^{T}(\vec{l}) = \frac{x - \mu_{ES_\alpha}}{\sigma_{ES_\alpha}} > \Phi^{-1}(1-\kappa)\,, \qquad (4.40)$$

where Φ denotes the CDF of the standard normal distribution. It should be noted that for a finite backtesting horizon, $Z_{ES_\alpha}^{T}$ is only approximately standard normally distributed under the assumption of a correctly specified risk measure forecast. This is in contrast to the VaR traffic light test, where observed violations follow a binomial distribution under a correctly specified model. Thus, for the VaR there is no need to employ an approximate normal distribution induced by the central limit theorem. Costanzino and Curran (2018) puts some further attention to the finite sample distribution of $Z_{ES_\alpha}^{T}$, but this goes beyond the scope of this thesis. Therefore, I rely on the assumption that critical values obtained by the standard normal distribution are reasonable approximations for a backtesting horizon like $T = 250$. In more detail, this means that critical values of $Z_{ES_\alpha}^{T,*} = 1.64$ and $Z_{ES_\alpha}^{T,*} = 3.72$ are going to be used within this thesis for the backtest significance levels of $\kappa = 0.05$ and $\kappa = 0.0001$ stipulated by the Basel Committee.

With respect to the criteria stated in Proposition 3.3, there is one major drawback of the ES backtest as it is proposed

[21] Note that Costanzino and Curran (2018) uses a different scaling for the ES violation rate compared to the original version by Costanzino and Curran (2015), which is also used within this thesis. This has no impact of the overall test methodology, but the achieved ES violation rates differ in their magnitude compared to Costanzino and Curran (2018).

by Costanzino and Curran (2015) and Costanzino and Curran (2018). Namely, for the calculation of the ES violation rate given in (4.35), which functions as a basis for the test decision, the risk manager needs to estimate the cumulative distribution function of returns $-L_t$ for any date within the backtesting horizon. This is especially problematic in case the risk manager applies a non-parametric ES forecasting model like a historical simulation. In such a situation, the backtest might not be feasible for the risk manager, as estimating the CDF of returns would be related to imposing further restrictions to the ES estimation model. Therefore, I will propose an approximate alternative to the backtest proposed by Costanzino and Curran (2015), which does not require the explicit derivation of any CDF function.

4.4.2. Adjusted approach - Approximate Z-test

Similar to the multinomial backtest proposed by Kratz et al. (2018), I suggest to approximate the integral $\int_\alpha^1 I_t(p)dp$ by a Riemann-sum using $N \in \mathbb{N}$ VaR figures with confidence levels between α and one. First of all, recall notation (4.2), i.e. $\alpha_j = \alpha + \frac{j-1}{N}(1-\alpha)$ for $j = 1,\dots,N$, for some reference confidence level α. Moreover, I propose to use the following approximation,

$$\int_\alpha^1 I_t(p)dp \approx \frac{1-\alpha}{N}\left(\sum_{j=1}^{N} I_t(\alpha_j)\right) . \tag{4.41}$$

This leads to an approximation of the ES violation ratio given in (4.35), which does not require any explicit estimation of a CDF function,

$$X_{ES_\alpha}^{T,\text{approx}} := \frac{1}{T(1-\alpha)}\sum_{t=1}^{T}\frac{(1-\alpha)}{N}\left(\sum_{j=1}^{N} I_t(\alpha_j)\right) , \tag{4.42}$$

for some $N \in \mathbb{N}$ which determines the approximation accuracy. It should be noted, that for the estimation of an realization $X_{ES_\alpha}^{T,\text{approx}}(\vec{l})$, one needs to estimate a vector of VaR figures, $(\widehat{VaR}_{t,\alpha_1},\dots,\widehat{VaR}_{t,\alpha_N})$. Nevertheless, in general this can be obtained from both parametric and non-parametric estimations models for the ES without any further difficulties. Furthermore, I define the approximative test statistic equivalent to (4.38) as,

$$Z_{ES_\alpha}^{T,\text{approx}} := \frac{X_{ES_\alpha}^{T,\text{approx}} - \mu_{ES_\alpha}}{\sigma_{ES_\alpha}} . \tag{4.43}$$

Additionally, the approximative approach also relies on the same one-sided test decision. The estimation model for the ES is rejected based on some return loss realizations $\vec{l}$, whenever

$$Z_{ES_\alpha}^{T,\text{approx}}(\vec{l}) > \Phi^{-1}(1-\kappa) , \tag{4.44}$$

where κ is again the significance level of the backtest. In order to limit the degree of misspecification due to an imprecise approximation, I propose a value of $N = 8$ for the calculation of $X_{ES_\alpha}^{T,\text{approx}}$ and $Z_{ES_\alpha}^{T,\text{approx}}$.

Overall, both versions presented within this section are indirect backtests for the ES. Whereas the original version from Costanzino and Curran (2015) requires the entire tail distribution of returns as an input parameter, the approximative version needs a vector of VaR estimates as an input variable. Similar as the multinomial approach by Kratz et al. (2018), both versions do not need actual ES forecasts as an input parameter. In case of a parametric ES estimation model, the original version brought up by Costanzino and Curran (2015) can easily be applied, given that the risk manager anyway has an explicit estimation of the return loss CDF at hand. On the other hand, for any non-parametric approach the approximative version suggested in the last subsection can be conducted, which although might be related to a higher degree of misspecification in the ES backtest. In the light of the further criteria stated in Proposition 3.3, both versions are computationally efficient and can easily be modified to a two-sided version. Moreover, I believe the underlying concept of generalizing the violation indicator for the VaR to the class of spectral risk measures is relatively intuitive. Furthermore, this allows for a high degree of flexibility as the approach can also be transferred to other spectral risk measures.

4.5. Residuals Bootstrap Test from McNeil and Frey (2000)

The *residuals bootstrap test* derived by McNeil and Frey (2000) is up to my knowledge the first backtest for the ES suggested within the literature. Nevertheless, it is still widely used as for example pointed out in Bayer and Dimitriadis (2019). Furthermore, it functions as a basis for different backtesting approaches suggested within the literature later on.[22] This subsection is first going to introduce the original version by McNeil and Frey (2000). Consecutively, a slight addition to the test decision will be made for the purpose of this thesis.

The test by McNeil and Frey (2000) is based on the so called *exceedance residuals*. In addition to VaR and ES forecasts, assume the risk manager also has a set of estimates for the conditional volatility of return losses at hand, denoted by $\{\widehat{\sigma}_t : t = 1,\dots,T\}$. For a set of realized return losses $\vec{l}$, the exceedance residual for any time t within the backtesting horizon is then given by,

$$r_t := \frac{l_t - \widehat{ES}_{t,\alpha}}{\widehat{\sigma}_t} I_t(\alpha) . \tag{4.45}$$

If one assumes that the applied risk estimation model is correctly specified, it holds

$$\mathbb{E}[r_t] = \mathbb{E}\left[\frac{L_t - ES_{t,\alpha}}{\sigma_t}\middle| L_t \geq VaR_{t,\alpha}\right] = 0 , \tag{4.46}$$

as the distribution of L_t is assumed to be continuous and strictly increasing, such that $ES_{t,\alpha} = \mathbb{E}[L_t | L_t \geq VaR_{t,\alpha}]$. On the contrary, if the applied estimation model underestimates the actual risk, i.e. $ES_{t,\alpha} > \widehat{ES}_{t,\alpha}$, then the expectation will be larger than zero. In order to estimate the expected value of exceedance residuals over the backtesting horizon, McNeil and Frey (2000)

[22] As an example the *Test 1* in Acerbi and Szekely (2014) can be seen as an adjusted version of the residuals bootstrap test proposed by McNeil and Frey (2000).

suggests to consider their mean value over the backtesting period given by,

$$\bar{r} := \frac{1}{\sum_{t=1}^{T} I_t(\alpha)} \sum_{t=1}^{T} r_t \, . \tag{4.47}$$

Thus, for the one-sided test version the following null and alternative hypothesis are going to be considered,

$$\begin{aligned} H0: &\quad \bar{r} \leq 0 \\ H1: &\quad \bar{r} > 0 \, . \end{aligned} \tag{4.48}$$

The actual test decision proposed by McNeil and Frey (2000) follows a bootstrap procedure, which is described in detail in Efron and Tibshirani (1994) on page 224. In short, the bootstrap based on a set of forecasts for VaR, ES and σ as well as realized return losses $\vec{l}$ is outlined below.

- Calculate the value of $\bar{r}$ related to the vector of observed losses and the risk estimates, i.e. calculate $\bar{r}(\vec{l})$.
- In order to simulate $\bar{r}$ under the $H0$ proceed in the following way. For any violation of $\widehat{VaR}_{t,\alpha}$, calculate the adjusted exceedance residual given by $r_t^{\text{adj}} = (r_t - \bar{r}(\vec{l}))\mathbb{1}_{\{r_t \neq 0\}}$.
- Set up the sample of adjusted residuals as $S := \{r_t^{\text{adj}} : r_t \neq 0\}$. If x observations of $\vec{l}$ breach the respective $\widehat{VaR}_{t,\alpha}$ level, then $|S| = x$, where $|S|$ denotes the cardinality of S. For each of the M bootstrap trials, draw $|S|$ observations with replacement from S, and calculate $\bar{r}^i$ for each trial $i \in \{1, \ldots, M\}$.
- Estimate the p-value related to the vector of observed loss $\vec{l}$ given by,

$$p = \frac{1}{M} \sum_{i=1}^{M} \mathbb{1}_{\{\bar{r}^i > \bar{r}(\vec{l})\}}. \tag{4.49}$$

Again the null hypothesis of the one-sided bootstrap test suggested by McNeil and Frey (2000) is rejected, whenever $p < \kappa$ for some predefined significance level κ for the backtest. For the purpose of this thesis, the respective bootstrap test decisions are going to be based on $M = 1000$ simulation trials in the subsequent chapters.

There are two aspects which are often criticized about the test proposed by McNeil and Frey (2000). First of all, the backtest requires an estimation of the volatility of return losses, which might not be feasible in certain scenarios. Indeed, as argued by Acerbi and Szekely (2014) and Bayer and Dimitriadis (2019) one can simply drop the volatility of return losses in definition (4.45) and use exceedance residuals which are not standardized by the volatility. Indeed, this can easily be implemented as it does not require any further adjustments of the overall test methodology. The second conceptional shortcoming of the backtest proposed by McNeil and Frey (2000), is that it only accounts for the magnitude of violations beyond the $VaR_{t,\alpha}$ threshold, but not for the overall amount of violations. As outlined in section 3.3 within this thesis, both would be necessary to account for a correct unconditional coverage of the ES. In other words, Acerbi and Szekely (2014) argues that the outlined test backtests the ES conditional on correctly specified VaR estimates. In order to handle this issue, I propose to apply the residual backtest from McNeil and Frey (2000) in combination with the VaR traffic light test introduced in section 3.2. Therefore, I propose the following adjusted one-sided null and alternative hypothesis.

$$\begin{aligned} H0: &\quad \bar{r} \leq 0 \quad \text{and} \quad \sum_{t=1}^{T} I_t(\alpha) \leq 1 - \alpha \\ H1: &\quad \bar{r} > 0 \quad \text{or} \quad \sum_{t=1}^{T} I_t(\alpha) > 1 - \alpha \end{aligned} \tag{4.50}$$

Correspondingly, the $H0$ in the combined backtest is rejected if either $p < \kappa$, where p is derived according to (4.49) or if $F^{binom}\left(\sum_{t=1}^{T} I_t(\alpha), T, 1-\alpha\right) > 1 - \kappa$. Therefore, I will use the *combined ES residuals backtest* based on the set of hypotheses (4.50) as it validates correct unconditional coverage of ES forecasts. Moreover, the test decision displays some kind of "worst-of-logic", as the null hypothesis is rejected, if either the original ES residuals backtest or the VaR traffic light test is rejected at a certain significance level κ.

With respect to the desirable properties of a backtest outlined in Proposition 3.3, the backtest described above is very intuitive from a conceptional point of view. Furthermore, it only requires realized P&L realizations as well as VaR and ES forecasts as input variables, given that the input of volatility forecasts is optional. The alleged shortcoming, that the backtest from McNeil and Frey (2000) only accounts for the magnitude but not the amount of VaR violations can easily be mitigated by linking the backtest to the VaR traffic light test. The only apparent disadvantage of the outlined test is, that the test decision again involves a simulation procedure.

Summing up this chapter, all five proposed ES backtest yield a one-sided test versions, in order to detect estimation models which underestimate the actual risk of an underlying portfolio. Furthermore, all introduced backtests are applicable for commonly used parametric and non-parametric ES estimation models, in a sense that all required data inputs can be obtained without inducing further assumptions. Overall, two of the five proposed tests are based on a simulation decision. Furthermore, for rather extreme backtest significance levels like $\kappa = 0.0001$, one should additionally also rely on the bootstrap version of Test 2 from Acerbi and Szekely (2014).

With respect to the practical aspects listed in Proposition 3.3, I propose the following qualitative judgement of all five overall testing methodologies based on the arguments outlined within this chapter. More precisely, I assign grades to all backtests in scope, with respect to the fulfilment of the four aspects listed in Proposition 3.3, namely *data intensity*, *computational effort*, *conceptual ease* and *clear decision making*. The grades are given by **(++)**, **(+)** for excellent or good results, **(o)** for an average result, and **(-)** or even **(- -)** if a certain aspect might be problematic for practical implementations. The proposed

judgement is outlined in Tables 3-7 below, together with a short explanation for any ES backtesting methodology.

- **Multinomial backtest - Original and de-facto one-sided version:**

All input variables for the multinomial backtesting methodology can easily be obtained for every parametric and non-parametric ES estimation approach. The backtest is computationally efficient as it does not rely on a bootstrap procedure and the methodology is a straightforward extension of the VaR traffic light test. Some adjustments to the test decision are needed in order to develop a one-sided test version, nevertheless this is not expected to have a major impact on the overall test performance.

- **Test 2:**

If the Test 2 is carried out with fixed critical values, it only requires VaR and ES forecasts in addition to P&L realizations, furthermore in this case no bootstrap is required. The overall methodology is rather intuitive and can also be adjusted to reflect a two-sided hypothesis. Especially for rather extreme backtesting confidence levels κ, an applicant faces the trade off between potentially inaccurate test decisions and a computationally costly bootstrap test version. Furthermore, in case a bootstrap is applied, additional input variables need to be stored.

- **Intercept ESR backtest:**

Input parameters for the intercept ESR backtest can be easily obtained for any ES estimation model, furthermore a standalone ES backtest version is proposed within Bayer and Dimitriadis (2019). On the contrary, the test methodology, especially the regression estimation based on the concept of conditional elicitabilty is rather complex. Test decisions can either be based on a Wald-test, which requires the cumbersome estimation of a consistent asymptotic covariance estimator as outlined in Bayer and Dimitriadis (2019), or on a bootstrap procedure, which is applied in the context of this thesis. Nevertheless, the bootstrap procedure is related to a rather extreme computational effort as a ML estimation needs to be conducted in every simulation trial. Both one and two-sided versions of the intercept ESR backtest can easily be conducted.

- **Z-Test - Original and approximative version:**

The original version requires the explicit estimation of the CDF of returns, which is not feasible for non-parametric ES forecasting models. This drawback can be avoided by applying the approximate version, which I suggested for the purpose of this thesis. Asymptotic distributional properties of the test statistic can be employed. Thus, no simulation procedure is required. Furthermore, the overall concept is rather intuitive. Both one-sided and two-sided test decisions can easily be obtained, nevertheless critical values for finite backtesting horizons are based on asymptotic distributional properties.

- **Combined ES residual backtest:**

The test methodology is very intuitive, furthermore only ES and VaR forecasts are required as mandatory input parameters in addition to P&L realizations. The test can be carried carried out as a one-sided or a two-sided version. Nevertheless, the test decisions is based on a simulation approach, which requires computational effort.

Overall, the concerns regarding the backtestability of the ES, which are for example stated in Carver (2013) and quoted within earlier stages of this thesis, appear to be exaggerated. Indeed, there are multiple promising backtest for the ES, some of which are listed above and are further evaluated in the consecutive chapters. With respect to practical applicability, especially the multinomial backtesting methodology as well as the combined ES residuals backtest need to be highlighted, as they exhibit the highest compliance with the criteria stated in Proposition 3.3. Nevertheless, all outlined ES backtesting versions qualify for a practical implementation.

5. Simulation study - Comparison of empirical size and power

This chapter aims to compare all five considered backtests as well as any adjusted version according to their empirical size and power. Within the literature both measures are commonly taken as a reference for the performance of statistical tests. For instance, analyses on both of these measures can be found in Acerbi and Szekely (2014), Bayer and Dimitriadis (2019) and Kratz et al. (2018), for the ES backtest presented within the respective article. On the other hand, up to the best of my knowledge, non of these two measures has been analysed yet for the Z-test introduced by Costanzino and Curran (2015). I will propose one single methodology for comparing simulated size and power of all ES backtests, which are in scope of this thesis. Thus, I will complement already existing analyses, in order to make a more founded judgement on the conceptual soundness of each of the evaluated approaches.

For the conducted simulation study, both versions of the multinomial backtest based on Kratz et al. (2018) as well as both version of the Z-test introduced by Costanzino and Curran (2015) are evaluated, in order to examine how the proposed adjustments impact both size and power of the original backtests. Furthermore, both test components of the combined ES residuals backtest are evaluated separately, to analyse how they contribute to the overall test decision. All other backtests are conducted as outlined in the preceding chapter.

The simulation study is structured as follows. Subsection 5.1 formally defines both size and power and introduces the set up of the analysis thereafter. Subsection 5.2 compares the size of all ES backtests, whereas subsection 5.3 compares their simulated power for different degrees of model misspecification. The chapter is concluded by a short judgement of the obtained size and power values for all backtests in scope.

5.1. Theory on empirical size and power

As for example outlined by Kratz et al. (2018), both size and power of an ES backtest are defined in the following way.

Data intensity	Computational effort	Conceptual ease	Clear decision making
(+)	(+ +)	(+ +)	(o)

Table 3: Assessment of multinomial ES backtest with respect to the criteria states in Proposition 3.3.

Data intensity	Computational effort	Conceptual ease	Clear decision making
(o)	(+)	(+)	(+)

Table 4: Assessment of the Test 2 with respect to the criteria states in Proposition 3.3.

Data intensity	Computational effort	Conceptual ease	Clear decision making
(+ +)	(-)	(-)	(+ +)

Table 5: Assessment of the intercept ESR backtest with respect to the criteria states in Proposition 3.3.

Data intensity	Computational effort	Conceptual ease	Clear decision making
(o)	(+ +)	(+)	(+)

Table 6: Assessment of the Z-test with respect to the criteria states in Proposition 3.3.

Data intensity	Computational effort	Conceptual ease	Clear decision making
(+)	(o)	(+ +)	(+ +)

Table 7: Assessment of the combined ES residuals backtest with respect to the criteria states in Proposition 3.3.

Definition 5.1 (Size/Power of ES backtest)**.** *Consider a backtest for the ES, then*

- *the size of the backtest is defined as* $\gamma := \mathbb{P}(\textit{reject } H0 | H0 \textit{ true})$. *Moreover, the size of the backtest corresponds to the respective type I error* γ.
- *the power of the backtest is defined as* $1-\beta := 1 - \mathbb{P}(\textit{accept } H0 | H0 \textit{ false})$. *Moreover, the power of the backtest corresponds to one minus the respective type II error* β.

On the one hand, any reasonable backtest for the ES should have a small size, which ought to be around the significance level κ applied in the backtesting procedure. As argued by Kratz et al. (2018), size values below κ are per se not problematic as they yield an even lower type I error. On the other hand, for a well functioning ES backtest, high power values, possibly close to one, are to be expected. Put in different words, a suitable ES backtest should yield a low probability of rejecting a correctly specified ES estimation model, while still being able to detect misspecified models with a high probability. Thus, in order to qualify for any practical implementation, an ES backtest should be as powerful as possible given a reasonably low size.

Within the literature, it is common practice to determine both size and power empirically by determining rejection rates in a simulation study. As mentioned above, respective simulation analyses can for example be found in Acerbi and Szekely (2014), Kratz et al. (2018) and Bayer and Dimitriadis (2019). The approach which will be applied within thesis is closest to the one outlined by Kratz et al. (2018). Sticking with the previously applied notation, a risk manager calculates an ES forecast based on some conditional return loss estimation $\widehat{L}_t$, whereas the true return loss is distributed according to L_t. Again both $\widehat{L}_t$ and L_t are based on the information available up to $t-1$, which is given by the σ-Algebra $\mathscr{F}_{t-1}$. The difference of the proposed setting compared to Kratz et al. (2018) is that time dependent risk forecasts are considered. This is necessary, as the methodology of the intercept ESR backtest from Bayer and Dimitriadis (2019) is based on the assumption of conditional, time dependent, risk forecasts. Thus, applying the intercept ESR backtest to a series of time-independent risk forecasts, ES_α for all $t \in \{1, \dots, T\}$, would not lead to any meaningful test decision.

For both size and power simulations different location-scale distributions, like the normal or the t-distribution, are applied for both $\widehat{L}_t$ and L_t, for observations t within the backtesting

horizon. Bayer and Dimitriadis (2019) uses a more sophisticated approach and fits ES estimation models to real financial data, nevertheless for both size and power simulations it is not essential that the assumed return loss distributions perfectly fit some underlying return series. Therefore, relying on commonly used location-scale distributions simplifies the set up of the applied analysis. In order to determine the size of the ES backtests, rejection rates over $MC = 1000$ test decisions are simulated in a scenario where $\widehat{L}_t = L_t$, i.e. the risk manager perfectly estimates the underlying risk. On the other hand, for the power calculations, rejection rates are based on $MC = 1000$ simulated test decisions in situations where $\widehat{L}_t \neq L_t$. More precisely, different scenarios are evaluated where the risk manager underestimates the actual risk of the underlying portfolio.

For the analysis conducted within this chapter, backtesting horizons of $T = 250$, 500 and 1000 observations are considered. Indeed, many authors argue that the backtesting period of 250 observations stipulated by the Basel Committee is not sufficient to backtest the ES. As an example, Kratz et al. (2018) considers a maximum backtesting period of $T = 2000$ observations, while Bayer and Dimitriadis (2019) even takes into account up to $T = 5000$ observations. Nevertheless, it should be noted that backtesting periods with $T > 1000$ are often not feasible in actual applications. Considering daily returns, 1000 observations correspond to a time series of roughly four years. Even in this case, it might be challenging for the risk manager to collect risk forecasts and return losses over such a long time frame. Furthermore, the conducted analysis takes into account a confidence level of $\alpha = 0.975$ for the ES and a significance level of $\kappa = 0.05$ for the backtests, which corresponds to the yellow traffic light zone in the methodology of the Basel Committee.

5.2. Empirical size of backtests

In order to determine the size of the backtest, I simulate $MC = 1000$ test decisions under the assumption that the risk manager correctly estimates the underlying risk, i.e. $\widehat{L}_t = L_t$ for all $t \in \{1,\dots,T\}$. Moreover, for every observation t within the backtesting horizon I set $\widehat{L}_t \sim N(\mu_t, \sigma_t)$ and $L_t \sim N(\mu_t, \sigma_t)$, respectively. For the procedure, at first 250 return loss observations $l_{-249},\dots,l_0$ are drawn from the standard normal distribution. Consecutively, both μ_1 and σ_1 are calculated as the sample mean and sample standard deviation of the preceding 250 observations. Furthermore, $\widehat{ES}_{1,\alpha}$ is calculated according to the theoretical value based on the distribution $N(\mu_1, \sigma_1)$. More precisely, the ES forecast based on the normal distribution is given by,

$$\widehat{ES}_{1,\alpha} = \frac{1}{1-\alpha}\sigma_1 \phi(\Phi^{-1}(\alpha)) + \mu_1 \,, \tag{5.1}$$

where again ϕ and Φ denote the PDF and the CDF of the standard normal distribution, respectively. Furthermore, also all auxiliary variables, needed for any of the outlined backtests, are calculated based on $N(\mu_1, \sigma_1)$. In the following, the actual observed loss is drawn from the same distribution, i.e $l_1 \sim N(\mu_1, \sigma_1)$. This procedure is afterwards carried out for any observation within the backtesting period, whereas both μ_t and σ_t are always calculated on the rolling sample of the previous 250 observations, i.e. $l_{t-250},\dots,l_{t-1}$. In the end, the decision of all evaluated backtests is based on the obtained vectors of $\{\widehat{ES}_{t,\alpha} : t = 1,\dots,T\}$ and $\{l_t : t = 1,\dots T\}$ as well as on potentially required auxiliary forecasts. Overall, $MC = 1000$ test decisions are simulated based on the outlined procedure and the respective rejection rate is calculated.

The resulting rejection rates for all considered backtests are depicted in Table 8. As previously mentioned, the ES is estimated at a confidence level of $\alpha = 0.975$. Moreover, a backtesting significance level of $\kappa = 0.05$ and backtesting horizons of $T = 250$, 500 and 1000 observations are considered. For visualisation purposes, the colouring scheme is taken from Kratz et al. (2018). Rejection rates in the size simulation below 6% yield good results and are thus coloured in green. Poor empirical size values are highlighted in orange if the simulated rejection rate exceeds 9% and in red given an even higher rejection rate above 12 %.

Overall, most of the evaluated backtests do show decent rejection rates in the size analysis over the considered backtesting horizons. Both multinomial backtests exhibit a rejection rate close to the significance level of 5% for all three backtesting time frames. As to be expected, the difference in rejection rates between both test versions is rather marginal. Furthermore, it is reasonable that rejection rates of the de-facto one-sided version are slightly below those of the original approach, as the adjusted backtest additionally rules out the rejection of simulation trials which display a conservative risk estimation. Overall, both multinomial backtests possess excellent size properties. The Test 2 from Acerbi and Szekely (2014) achieves low rejection rates, which are even below the significance level applied in the simulation study. Interestingly, it appears that rejection rates even converge to zero given an increase in the backtesting horizon. As argued before, this should per se not be seen as a drawback of the test from Acerbi and Szekely (2014). Nevertheless, if the Test 2 reveals a low power in the consecutive subsection, this might be an indicator for a poor balancing of both size and power in the respective test framework. The simulation results for the intercept ESR backtest are in line with expectations, with rejection rates below 7% in all three simulations. Furthermore, the empirical size of the intercept ESR backtest appears to be rather stable across different backtesting horizons, close to the applied significance level κ. On the contrary, the empirical size of both, the original Z-test from Costanzino and Curran (2015) as well as the proposed approximative version, exceeds a value of 10% for all evaluated scenarios. It is even more problematic that the simulated rejection rates further increase given an increase in the backtesting period. For a value of $T = 1000$ the original Z-test rejects the correctly specified ES estimation model in almost every fourth test decision. Surprisingly, the empirical size of the approximative Z-test seems to be slightly superior compared to the original version. Nevertheless, both test versions exhibit unsatisfactory high rejection rates in the size analysis, which hints towards a potential misspecification in the underlying framework of the ES backtest. On possible reason might be, that critical values based on the asymptotic distribution of the test statistic might not lead to accurate test decisions for finite backtesting horizons like $T = 250$, as shortly

Empirical size - Rejection rates of evaluated ES backtests

	Backtesting period - T		
ES backtesting approach	250	500	1000
Multinomial backtest - Original version	**0.040**	**0.040**	**0.051**
Multinomial backtest - One-sided version	**0.037**	**0.040**	**0.050**
Test 2	**0.042**	**0.012**	**0.000**
Intercept ESR backtest	0.063	**0.043**	0.063
Z-test - Original version	**0.117**	**0.171**	**0.240**
Z-test - Approximative version	**0.101**	**0.117**	**0.171**
Combined ES residuals backtest	**0.098**	**0.110**	0.085
(ES residuals backtest)	**(0.056)**	**(0.030)**	**(0.024)**
(VaR traffic light test)	**(0.044)**	(0.083)	(0.062)

Table 8: Empirical size - rejection rates of evaluated ES backtests. ES confidence level set at $\alpha = 0.975$, significance level of the backtests at $\kappa = 0.05$. Rejection rates are based on $MC = 1000$ simulations given that $\widehat{L}_t = L_t \sim N(\mu_t, \sigma_t)$ for all $t \in \{1, \ldots, T\}$.

outlined in chapter 4.4. The combined ES residuals backtest also shows slightly too high rejection rates ranging between 8% and 11%. It should be noted, that the test decision of the combined test is composed of both the VaR traffic light test as well as the original ES residuals backtest proposed by McNeil and Frey (2000). Both of these two single test decisions show reasonable size values in all three simulation scenarios. As outlined in chapter 4.5, the combined ES residual backtest is rejected if any of the two applied test components is rejected. Therefore, given the worst-of-logic applied for the combined test decision, it is not surprising that the combined ES residuals backtest is slightly oversized. One might decrease the significance levels for each of the two single test components in order to calibrate the combined ES residuals backtests to a size value closer to 5%. Given that the size values of the combined test are still within an acceptable range, no rescaling is applied and the backtest will be carried out as described in the previous chapter.

Concluding, both multinomial backtests based on Kratz et al. (2018) and the intercept ESR backtest from Bayer and Dimitriadis (2019) show decent size properties in line with expectations. Size values close to zero, observed for the Test 2 from Acerbi and Szekely (2014), might be seen as a bonus of the test framework, if the test still reveals decent power values. Nevertheless, this needs to be evaluated in the following subsection, as overly low size figures might also go hand in hand with a low power in detecting misspecified ES estimation models. The combined ES residuals backtest is slightly oversized. Nevertheless, this can be explained given the underlying test methodology and furthermore one might mitigate too high rejection rates by calibrating the single components of the test decision. Only the size values of the Z-test proposed by Costanzino and Curran (2015) deviate from expectations. It might be difficult for regulators to rely on an ES backtest which rejects a correctly specified ES estimation model in up to every fourth scenario depending on the selected testing parameters.

Furthermore, also no clear pattern can be observed across all backtests regarding the relation of empirical size values and the choice of the backtesting horizon. Thus, it is not obvious that an increased backtesting period automatically leads to improved size properties based on the conducted simulations.

5.3. Empirical power of backtests

This subsection aims to evaluate the empirical power of all considered ES backtests, i.e. their ability to detect certain levels of model misspecification. As I decided to base my analysis on one-sided test decisions, only scenarios are analysed where the risk manager underestimates the true underlying risk given by L_t. For all single power scenarios, again 250 initial loss observations are drawn from the standard normal distribution. Again both μ_t and σ_t are estimated over the previous 250 observations for every observation t within the backtesting horizon. Furthermore, the risk manager again estimates the underlying risk $\widehat{ES}_{t,\alpha}$ based on the respective normal distribution, i.e. $\widehat{L}_t \sim N(\mu_t, \sigma_t)$. The difference compared to the previous size analysis is that the true distribution L_t deviates from the normal distribution. More precisely, observed return losses l_t, for $t \in \{1, \ldots, T\}$, are drawn from different location-scale distributions, whereas location and scale parameters again depend on both μ_t and σ_t. Moreover, four different scenarios are analysed, which can be grouped into the following two categories of model misspecification.

(i) **Misspecified tail behaviour and conditional variance**
For the first category, I assume that the true return loss follows a t-distribution with location parameter μ_t and scale parameter σ_t. Moreover, I consider two scenarios with $\nu = 3$ and $\nu = 5$ degrees of freedom. Thus, the observed return losses l_t are drawn from $L_t \sim t_\nu(\mu_t, \sigma_t)$. As the t-distribution possesses fatter tails compared to the normal distribution, this leads to a scenario where the risk manager underestimates the underlying risk, due to a misspecification of the distribution tail. The tail

behaviour of the t-distribution converges to that of the normal distribution for increasing values of ν. Therefore, the degree of risk underestimation is larger for the case of $\nu = 3$ compared to $\nu = 5$.
Secondly, the variance of $L_t \sim t_\nu(\mu_t, \sigma_t)$ is given by $Var(L_t) = \sigma_t^2 \frac{\nu}{\nu-2}$, whenever $\nu > 2$. Taking into account that the risk manager assumes return losses to be normally distributed, this leads to

$$Var(L_t) = \sigma_t^2 \frac{\nu}{\nu-2} > \sigma_t^2 = Var(\widehat{L}_t), \quad \text{for } \nu > 2. \tag{5.2}$$

Therefore, the risk manager additionally also underestimates the true conditional variance of return losses. Again, the degree of risk underestimation due to a misspecified variance is larger for the case of $\nu = 3$ compared to $\nu = 5$.
As previously argued, both the number of violations beyond the respective VaR threshold as well as their magnitude with respect to the estimated ES value, need to be considered for correct unconditional coverage of the ES. The misspecified tail behaviour will become apparent in the outer tail of the return loss distribution and thus primarily impact the magnitude of observed VaR violations. Additionally, due to the misspecified conditional variance, the overall level of riskiness is underestimated, which will result in a rather high number of violations beyond the forecasted VaR threshold. Therefore, for scenarios from the first category, not just ES estimates, but also the related VaR forecasts, are inaccurate.
For illustration purposes, Figure 1 depicts the PDF of the forecasted return loss distribution $\widehat{L}_t$ as well as the PDFs of both considered t-distributions with parameter values of $\mu_t = 0$ and $\sigma_t = 1$.

In order to complement the two selected scenarios, I additionally analyse how rejection rates evolve given that the number of considered degrees of freedom further increases beyond a value of $\nu = 5$.

(ii) **Misspecified tail behaviour**
For the second category, I analyse two scenarios where the risk manager only fails to correctly account for the distribution tail of return losses. Nevertheless, compared to the first category I evaluate scenarios where the risk manager correctly models both the conditional mean and variance of the underlying distributions. Thus, the level of risk underestimation in the second category is lower compared to the first category. More precisely, the underestimation of the actual risk will primarily become apparent in the magnitude of violations beyond the estimated VaR but not in their overall number. Thus, the foreasted VaR, for scenarios from the second category, might be roughly accurate, whereas the related ES figure still underestimates the actual underlying risk. Therefore, this makes it also more challenging for any ES backtest to detect a misspecified model, which stems from this second category.
For the first scenario in this category, I simulate observed return losses from a t-distribution with $\nu = 3$ degrees of freedom, which is standardized to have a conditional mean of $\mathbb{E}[L_t] = \mu_t$ and more importantly a conditional variance of $Var(L_t) = \sigma_t^2$. This is achieved by simulating observed return losses from $L_t \sim t_3(\mu_t, \sigma_t/\sqrt{3})$. Correspondingly, the risk manager takes into account the correct conditional variance, but does not account for the fat tails of the t-distribution.
For the second scenario in this category, I simulate return losses from a *skewed normal distribution*, which is again standardized such that both conditional mean and variance are correctly specified. In addition to location and scale parameters, the skewed normal distribution requires an additional shape input parameter λ to determine the skewness of the distribution. The properties of the skewed normal distribution are for example outlined in Azzalini and Valle (1996). Given location and scale parameters of μ and σ and a shape parameter of λ, the PDF of the skewed normal distribution is given by,

$$\phi^{\text{skewed}}(x) := \frac{2}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)\Phi\left(\lambda\frac{x-\mu}{\sigma}\right), \tag{5.3}$$

where again ϕ and Φ denote the PDF and the CDF of the standard normal distribution. Moreover, I select a shape parameter of $\lambda = 2$ and standardize both the location and the scale parameter such that $\mathbb{E}[L_t] = \mu_t$ and $Var(L_t) = \sigma_t^2$.[23] The resulting return loss distribution of L_t is right-skewed, which is a realistic property of a financial time series of return losses. Furthermore, the distribution possesses over-kurtosis compared to the standard normal distribution. Thus, the risk manager underestimates the underlying risk, as he fails to correctly estimate both skewness and kurtosis of the return loss distribution. Figure 2 depicts the PDF of the estimated return loss $\widehat{L}_t$ as well as the PDFs of both assumed data generating processes within this category. Again for illustration purposes, location and scale parameters of $\mu_t = 0$ and $\sigma_t = 1$ are considered.

At first, the more severe cases of risk underestimation are evaluated, where both tail behaviour and conditional variance are wrongly estimated. In the first power analysis, observed losses are drawn from $L_t \sim t_3(\mu_t, \sigma_t)$, while the risk manager bases his ES forecasts on the estimate $\widehat{L}_t \sim N(\mu_t, \sigma_t)$. The respective simulated rejection rates are depicted in Table 9. The colour scheme for the power analysis is again motivated by Kratz et al. (2018). Nevertheless, the respective thresholds are taken at stricter values compared to their suggestion. Power values are marked in green given a high rejection rate above 80 %. On the other hand, poor power values are marked in orange given a

[23] For $\lambda = 2$ this is achieved by setting the scale parameter of the skewed normal distribution equal to $\frac{\sigma_t}{\sqrt{1-1.6/\pi}}$ and the location parameter equal to $\mu_t - \frac{\sigma_t}{\sqrt{1-1.6/\pi}}\frac{2}{\sqrt{5}}\sqrt{\frac{2}{\pi}}$.

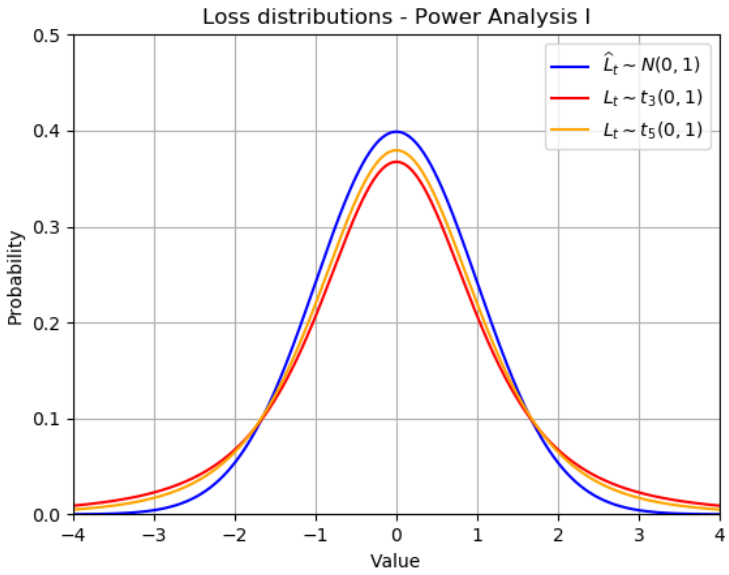


Figure 1: Power analysis - Comparison of PDFs corresponding to multiple loss distributions L_t to misspecified distribution $\widehat{L}_t$ with different variance and tail behaviour.

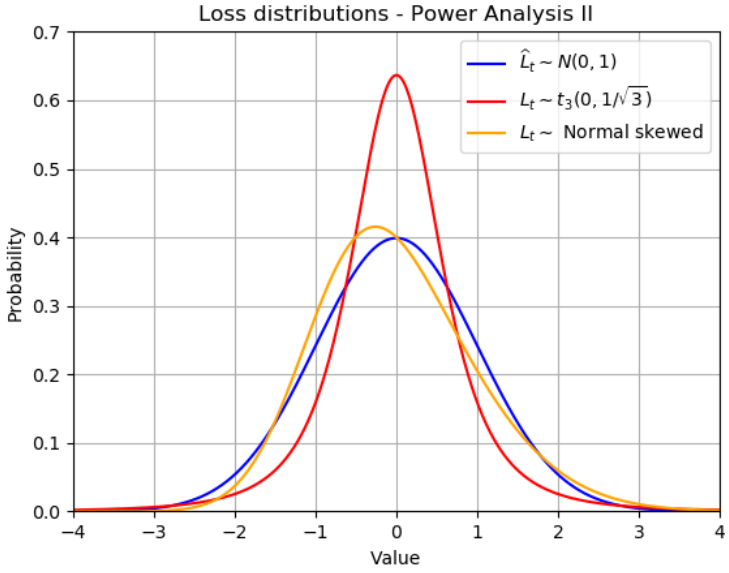


Figure 2: Power analysis - Comparison of PDFs corresponding to multiple loss distributions L_t to misspecified distribution $\widehat{L}_t$ with different tail behaviour but same variance.

rejection rate below 40 % and in red if the simulated rejection rate even lies below 20 %. For the power analysis, rejection rates are again based on $MC = 1000$ simulated test decisions. Furthermore, an ES confidence level of $\alpha = 0.975$ and a backtest significance level of $\kappa = 0.05$ are applied for backtesting periods of $T = 250$, 500 and 1000 observations.

First of all, it can be noticed that all ES backtests in scope yield excellent power values in Table 9, as all simulated rejection rates are above 78 %. There is only one rejection rate that is not labelled in green, however only with a value slightly below 80 %. Moreover, this rejection rate belongs to the intercept ESR backtest in case of $T = 1000$ backtesting observations. It is also worth mentioning that rejection rates of the intercept ESR backtest decrease given an increase in the backtesting horizon T, which is rather counter-intuitive. Nevertheless, the intercept ESR backtest still exhibits decent power values in all evaluated cases. On the contrary, rejection rates for all other considered ES backtests increase in T. Furthermore, for both $T = 500$ and

Empirical power $L_t \sim t_3(\mu_t, \sigma_t)$ - Rejection rates of evaluated ES backtests

	Backtesting period - T		
ES backtesting approach	250	500	1000
Multinomial backtest - Original version	**0.988**	**1.000**	**1.000**
Multinomial backtest - One-sided version	**0.988**	**1.000**	**1.000**
Test 2	**0.992**	**1.000**	**1.000**
Intercept ESR backtest	**0.914**	**0.888**	0.789
Z-test - Original version	**0.950**	**0.975**	**0.990**
Z-test - Approximative version	**0.995**	**1.000**	**1.000**
Combined ES residuals backtest	**0.987**	**1.000**	**1.000**
(ES residuals backtest)	**(0.881)**	**(0.975)**	**(0.997)**
(VaR traffic light test)	**(0.971)**	**(1.000)**	**(1.000)**

Table 9: Empirical power, i.e. rejection rates of evaluated ES backtests. Rejection rates are based on $MC = 1000$ simulations given that $\widehat{L}_t \sim N(\mu_t, \sigma_t)$ and $L_t \sim t_3(\mu_t, \sigma_t)$ for all $t \in \{1, \dots, T\}$. ES confidence level set at $\alpha = 0.975$, significance level of the backtests at $\kappa = 0.05$. Backtesting horizons of $T = 250$, 500 and 1000 observations are considered.

$T = 1000$ observations, a majority of the backtests in scope reject the wrongly specified risk forecast at a perfect rate of 100%. It is in line with expectations, that power values increase given an increase in T, as a larger backtesting period also involves more misspecified risk forecasts into the respective test decision.

For the multinomial backtest there is no difference in rejection rates between the original and the de-facto one-sided version. As outlined in the previous chapter, the only difference between both multinomial tests is, that the adjusted approach additionally rules out the rejection of any conservative estimation model. Nevertheless, for the power analysis, only scenarios where the actual risk is underestimated are considered. Therefore, it is reasonable that rejection rates of both multinomial test versions coincide for any power scenario. For the Z-test proposed by Costanzino and Curran (2015) it can be noted that the approximative version, proposed within this thesis, yields slightly higher rejection rates compared to the original approach. Overall, all evaluated backtests show decent results in detecting the risk underestimation in a scenario where $\widehat{L}_t \sim N(\mu_t, \sigma_t)$ and $L_t \sim t_3(\mu_t, \sigma_t)$.

In order to stepwise increase the challenge, I secondly consider a scenario where $L_t \sim t_5(\mu_t, \sigma_t)$. As outlined above, increasing the degrees of freedom decreases the degree of misspecification of both the distribution tail and the conditional variance. Thus, it is more difficult to detect, that the risk manager still underestimates the actual risk. The respective rejection rates for this scenario are depicted in Table 10 below. Both the colour scheme and the selected backtesting parameters are equivalent to the previous table.

Similar as in the previous scenario, all considered ES backtests exhibit excellent rejection rates, most of them again above 80 %. Nevertheless, the overall level of rejection rates is slightly lower compared to the previous simulation, which is in line with expectations given the decreased level of model misspecification. As an example, for the simulation of $L_t \sim t_3(\mu_t, \sigma_t)$, multiple backtests achieved a perfect rejection rate already at a backtesting horizon of $T = 500$ observations, whereas for this second scenario, using $L_t \sim t_5(\mu_t, \sigma_t)$, perfect rejection rates of 100 % are only achieved for the maximum considered backtesting time frame of $T = 1000$ observation. Thus, a risk manager needs to consider a longer time series of both risk forecasts and P&L realizations, to achieve the same backtesting accuracy as in the previous scenario. Furthermore, all ES backtests analysed in Table 10 generally, show increasing rejection rates given an increase in the backtesting horizon T. One can also notice, that the original Z-test proposed by Costanzino and Curran (2015) is again slightly out-performed by its approximative version, which was suggested within this thesis.

Although there are some minor differences between the ES backtests in scope, all of them show excellent empirical power values in both considered scenarios, where the risk manager misspecifies both the tail behaviour and the conditional variance. This is also in line with previous power analyses conducted for example by Acerbi and Szekely (2014). In order to evaluate how rejection rates further develop if the true return loss distribution approaches the normal distribution, I outline some additional analysis in the following. Moreover, I simulate $MC = 100$ test decisions for each scenario $L_t \sim t_\nu(\mu_t, \sigma_t)$, whereas ν takes on values in $\{3, \dots, 25\}$, and the estimated return loss $\widehat{L}_t$ is again assumed to be normally distributed. Furthermore, I choose the same backtesting parameters as in the previous scenarios, but fix a backtesting horizon of $T = 500$ observations. Thus, rejection rates are simulated for the same type of model misspecification as outlined in the previous two scenarios, but with a decreasing level of estimation inaccuracy. The resulting plot of rejection rates, depending on the considered degrees of freedom, for all considered backtests, is depicted in Figure 3[24] .

[24] Note that both the original two-sided and the de-facto one sided multino-

Empirical power $L_t \sim t_5(\mu_t, \sigma_t)$ - Rejection rates of evaluated ES backtests

	Backtesting period - T		
ES backtesting approach	250	500	1000
Multinomial backtest - Original version	**0.836**	**0.956**	**1.000**
Multinomial backtest - One-sided version	**0.836**	**0.956**	**1.000**
Test 2	**0.900**	**0.957**	**0.995**
Intercept ESR backtest	**0.847**	**0.963**	**0.995**
Z-test - Original version	**0.817**	**0.862**	**0.884**
Z-test - Approximative version	**0.950**	**0.995**	**1.000**
Combined ES residuals backtest	**0.895**	**0.992**	**1.000**
(ES residuals backtest)	(0.655)	**(0.895)**	**(0.994)**
(VaR traffic light test)	**(0.801)**	**(0.971)**	**(1.000)**

Table 10: Empirical power, i.e. rejection rates of evaluated ES backtests. Rejection rates are based on $MC = 1000$ simulations given that $\widehat{L}_t \sim N(\mu_t, \sigma_t)$ and $L_t \sim t_5(\mu_t, \sigma_t)$ for all $t \in \{1, \ldots, T\}$. ES confidence level set at $\alpha = 0.975$, significance level of the backtests at $\kappa = 0.05$. Backtesting horizons of $T = 250$, 500 and 1000 observations are considered.

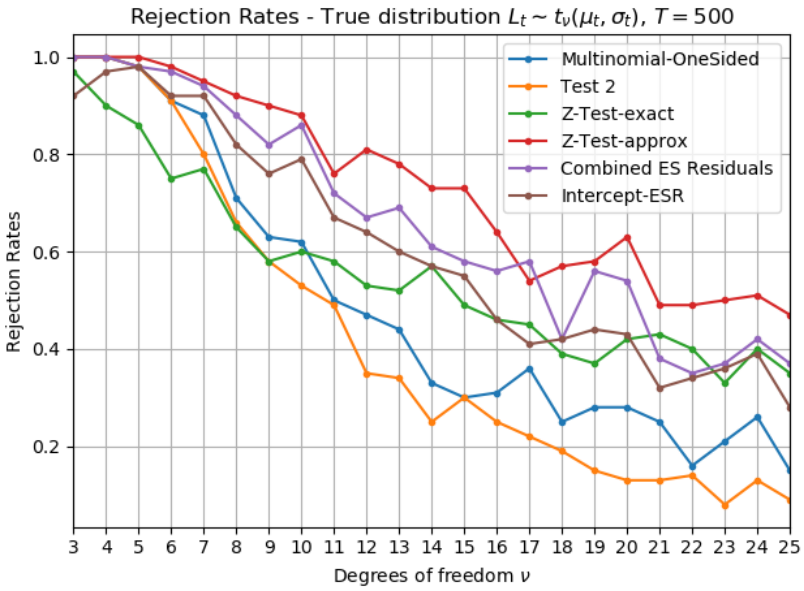


Figure 3: Rejection rates for all considered ES backtests depending on degree of model misspecification. Moreover, $MC = 100$ test decisions are simulated for each scenario $\widehat{L}_t \sim N(\mu_t, \sigma_t)$ and $L_t \sim t_\nu(\mu_t, \sigma_t)$, whereas ν takes on values in $\{3, \ldots, 25\}$. ES confidence level set at $\alpha = 0.975$, significance level of the backtest set at $\kappa = 0.05$, based on a backtesting horizon of $T = 500$ observations.

Overall, the results depicted in Figure 3 are in line with expectation, as rejection rates decrease given an increase in degrees of freedom. Furthermore, up to a value of $\nu = 10$ all considered ES backtests reject the underestimated ES forecast in more than every second test decisions. For larger values of ν around 20, both applied Z-tests, the combined ES residuals backtest as well as the intercept ESR backtest display slightly higher rejection rates compared to both, the multinomial backtest as well as the Test 2. Nevertheless, for high values of ν, minor differences between the single backtests should not be exaggerated, given the rather low degree of model inaccuracy. Overall, the power of all ES backtests in scope is deemed appropriate for scenarios from the first category of model misspecification.

In the following the two selected scenarios from the second category are analysed, where only the risk in the distribution tail is underestimated, but the conditional first two moments are correctly specified, i.e. $\mathbb{E}[\widehat{L}_t] = \mathbb{E}[L_t]$ and $Var(\widehat{L}_t) = Var(L_t)$ for

mial backtest achieve exactly the same rejection rates in the conducted analysis. Thus, only the one-sided version is included in Figure 3.

all observations t within the backtesting period. Thus, the underestimation of the actual risk only becomes perceivable in the outer tail of the return loss distribution. As previously argued, this primarily affects the magnitude of violations beyond the respective VaR threshold, but not there overall number. Therefore, it is more challenging for any ES backtest to detect a misspecified model from this second considered category.

Table 11 depicts the simulated rejection rates for the scenario where the true return loss variable follows a standardized t-distribution with $\nu = 3$ degrees of freedom.

Again rejection rates are based on $MC = 1000$ simulated test decisions given an ES confidence level of $\alpha = 0.975$ and a significance level of $\kappa = 0.05$ for the considered backtests. The colouring scheme is equivalent to the previous tables.

Overall, it can be recognized that power values are by far worse given that L_t follows a standardized t-distribution compared to the previous simulations, where the variance of the t-distribution was not adjusted. For the backtesting period of $T = 250$ days, stipulated by the Basel Committee, non of the evaluated backtests manages to reject the misspecified ES estimation model in at least half of all simulated decisions. For most backtests in scope, rejection rates tend to increase given an increase in the considered backtesting horizon.

Nevertheless, at least for two out of the five overall testing methodologies, the power results are not even satisfying for the maximum backtesting period of $T = 1000$ days. Both, the original Z-test proposed by Costanzino and Curran (2015) as well as the related approximative version, reject the underestimated ES estimate in less than one third of all test decisions for any considered backtesting period. Although rejection rates for both Z-tests slightly increase in T, both tests are not able to consistently detect this type of model misspecification. Furthermore, results obtained for the Test 2 from Acerbi and Szekely (2014) are even more problematic, as all simulated rejection rates for this test are labelled in red, with values even below 20 %. The rejection rates even further decrease given an increase in the backtesting time frame. Thus, the Test 2 does not appear to have any power to detect this kind of risk underestimation. Taking into account the size simulations in the previous subsection, the Test 2 exhibits a potential imbalanced relation between both size and power. Indeed, extremely low size values might come at the cost of a low power in detecting certain types of model misspecification.

On the contrary, at least for the maximum considered backtesting period, both multinomial backtest versions as well as the intercept ESR backtest achieve rejection rates above 50 %. Moreover, the rejection rate for the combined ES residuals backtest is even labelled in green with an excellent value of 83.7 %, given $T = 1000$ backtesting observations are taken into account. As the risk manager correctly estimates the conditional variance of return losses, it is not surprising that the VaR traffic light test produces rejection rates close to zero, as the misspecification only gets perceivable in the outer distribution tail. Still, in combination with the original ES residuals backtest proposed by McNeil and Frey (2000), the combined test shows the best empirical power values amongst all evaluated approaches in this simulation scenario.

For the last considered power simulation, true return losses are simulated from a skewed normal distribution, which is standardized such that both conditional mean and variance coincide with the respective figures of the estimated return loss variable $\widehat{L}_t$. Whereas $\widehat{L}_t$ models a symmetric return loss distribution, L_t is right-skewed and also possesses a slightly higher kurtosis. The simulated rejection rates are depicted in Table 12 below. All backtesting parameters are set equivalent to the previous table.

First of all, it can be noticed that power figures are on a slightly higher level compared to the previous scenario. This is in line with expectations, as the true return loss distribution L_t is both skewed and fat tailed, which induces a slightly higher degree of misspecification in the distribution tails compared to the previous simulation. For this last simulated power scenario, it becomes most apparent, that an increasing backtesting horizon is related to higher empirical power values. While most simulated rejection rates do not exceed 50 % for $T = 250$ backtesting observation, the majority of the backtesting approaches in scope does a decent job in detecting the misspecified model in case a backtesting horizon of $T = 1000$ is considered. Especially, the intercept ESR backtest, the approximative Z-test as well as the combined ES residuals backtest need to be highlighted with excellent rejection rates above 90 %, given the maximum considered backtesting period. As an exception, again the Test 2 from Acerbi and Szekely (2014) is not able to consistently detect this kind of risk underestimation. Indeed, for any of the considered backtesting time frames, the Test 2 detects the wrongly specified estimation model only in about 40 % of all simulated test decisions.

Concluding this subsection, all considered backtests in scope achieve excellent empirical power values in scenarios where the risk manager underestimates both the conditional variance as well as the distribution tail of return losses. On the contrary, it is far more difficult for any ES backtest to detect scenarios where the first two moments of return losses are correctly estimated but only the distribution tail is underestimated. Nevertheless, for these last two scenarios from the second category, differences between the considered ES backtests regarding their empirical power become most apparent. The Test 2 from Acerbi and Szekely (2014) performs worst in these two simulations, as it only exhibits few power in detecting the respective misspecified estimation models. As previously argued, this might be related to an imbalanced relation of both empirical size and power observed for the Test 2. The most consistent power values across both scenarios form the second category are achieved by, both multinomial backtests, the intercept ESR backtest as well as the combined ES residuals backtest, with decent rejection rates, at least for the maximum backtesting horizon of $T = 1000$ observations. Overall, it should also be noted that the approximative Z-test slightly outperforms the original test version proposed by Costanzino and Curran (2015). Although this appears to be surprising, this trend can be observed throughout all four conducted power simulations. Furthermore, in line with expectations, rejection rates of both multinomial test version coincide in all conducted power analyses.

Compared to the size analysis in the previous subsection, one

Empirical power $L_t \sim t_3(\mu_t, \sigma_t/\sqrt{3})$
- Rejection rates of evaluated ES backtests

	Backtesting period - T		
ES backtesting approach	250	500	1000
Multinomial backtest - Original version	**0.149**	**0.287**	0.616
Multinomial backtest - One-sided version	**0.149**	**0.287**	0.616
Test 2	**0.152**	**0.080**	**0.028**
Intercept ESR backtest	**0.207**	**0.273**	0.509
Z-test - Original version	**0.176**	**0.231**	**0.293**
Z-test - Approximative version	**0.204**	**0.233**	**0.324**
Combined ES residuals backtest	0.414	0.547	**0.837**
(ES residuals backtest)	(0.408)	(0.543)	**(0.837)**
(VaR traffic light test)	**(0.018)**	**(0.028)**	**(0.015)**

Table 11: Empirical power, true dist. t_3 standardized - rejection rates of evaluated ES backtests. ES confidence level set at $\alpha = 0.975$, significance level of the backtests at $\kappa = 0.05$. Rejection rates are based on $MC = 1000$ simulations given that $\widehat{L}_t \sim N(\mu_t, \sigma_t)$ and $L_t \sim t_3(\mu_t, \sigma_t/\sqrt{3})$ for all $t \in \{1, \ldots, T\}$.

Empirical power $L_t \sim$ skewed normal -
Rejection rates of evaluated ES backtests

	Backtesting period - T		
ES backtesting approach	250	500	1000
Multinomial backtest - Original version	**0.308**	0.488	0.755
Multinomial backtest - One-sided version	**0.308**	0.488	0.755
Test 2	**0.396**	**0.388**	0.410
Intercept ESR backtest	0.467	0.701	**0.906**
Z-test - Original version	0.442	0.594	0.603
Z-test - Approximative version	0.594	**0.833**	**0.968**
Combined ES residuals backtest	0.457	0.747	**0.927**
(ES residuals backtest)	**(0.179)**	**(0.358)**	(0.625)
(VaR traffic light test)	**(0.349)**	(0.633)	**(0.821)**

Table 12: Empirical power, true distribution skewed normal - rejection rates of evaluated ES backtests. ES confidence level set at $\alpha = 0.975$, significance level of the backtests at $\kappa = 0.05$. Rejection rates are based on $MC = 1000$ simulations given that $\widehat{L}_t \sim N(\mu_t, \sigma_t)$ and L_t follows a skewed normal distribution, which is standardized such that first two moments of $\widehat{L}_t$ and L_t coincide for all $t \in \{1, \ldots, T\}$.

can also observe an obvious impact of the backtesting time frame on power values. Especially, for both scenarios from the second category ES backtests are generally more powerful given a larger backtesting horizon. Moreover, for the last two scenarios, the backtesting period of $T = 250$ observations, suggested by the Basel Committee, does not appear to be sufficient in order to consistently detect misspecified ES estimation models.

Summing up this chapter, I propose the following qualitative judgement of all ES backtesting methodologies in scope according to their results in all conducted size and power simulations. All backtests are graded according to both their empirical size and power values. In each category **(++)** corresponds to an excellent performance, followed by **(+)**, which still characterises an above average results. Average power or size results are labelled by **(o)**, whereas poor performance is marked with a **(-)** or even with a **(- -)** if it displays a serious issue for any practical implementation. The judgement of the backtests in scope as well as a short explanation is outlined below.

- **Multinomial backtest - Original and de-facto one-sided version :**
  Size: (++) Excellent empirical size for both the original and the de-facto one-sided version. Furthermore, the de-facto one-sided version avoids the rejection of any

conservative ES estimation model.
Power: (+) Excellent empirical power for scenarios from first category. Still decent power for scenarios from second category, given a sufficient backtesting horizon. As to be expected, there are no differences in power values between both multinomial test versions.

- **Test 2:**
  Size: (+) Very low empirical size, which is slightly below backtesting significance level and further decreases given an increase in T.
  Power: (-) Excellent empirical power for scenarios from first category. On the contrary, low power in detecting misspecified models from the second category. Overall, the obtained results hint towards a slight misalignment of both size and power.

- **Intercept ESR backtest:**
  Size: (+) Decent empirical size, which is nevertheless slightly above backtesting significance level.
  Power: (+) Excellent empirical power for scenarios from first category. Still decent power for scenarios from second category, given a sufficient backtesting horizon.

- **Z-test - Original and approximative version:**
  Size: (- -) Both the original as well as the approximate version are oversized. Furthermore, empirical size further increases with increasing values of T. High size values indicate a potential misspecification in the underlying test framework.
  Power: (+) Excellent power for scenarios from first category, but both versions can not consistently detect misspecified models from second category. Overall, approximative version yields slightly superior power results compared to the original approach.

- **Combined ES residuals backtest:**
  Size: (o) Backtest is slightly oversized, but might be adjusted by calibration of single test components.
  Power: (++) Excellent empirical power for scenarios from first category and also for the scenarios from the second category, given a sufficient backtesting horizon T.

Therefore, according to both empirical size and power, the de-facto one-sided version of the multinomial backtest yields the best results followed by both the intercept ESR backtest and the combined ES residuals backtest.
In the consecutive chapter, all backtests in scope are tested on reasonable ES estimation models, which are fitted to log-return losses of the S&P 500.

6. Application of backtests to real data

The object of this chapter is to apply the relevant ES backtests outlined in the fourth chapter of this thesis to actual financial time series. Thus, this chapter aims to evaluate whether the ES backtests in scope yield reasonable results in practical scenarios and therefore qualify for the use by both financial institutions and regulators. For this purpose, univariate price data of the S&P 500 index is considered over two distinct time periods, which function as the respective backtesting horizons in this context. The consideration of two different time periods of S&P 500 data allows for an evaluation of the introduced ES backtests in different market environments. The evaluated data set is further described in the consecutive subsection. Afterwards, five ES estimation models, with different degrees of complexity and forecasting accuracy, are estimated for every day within each of the two considered time frames of S&P 500 data. The respective risk models are further outlined in subsection 6.2. In the subsection thereafter, the estimated ES forecasts are evaluated based on all ES backtesting methodologies introduced in the course of this thesis. Similar as for the VaR traffic light approach, backtesting significance levels of both $\kappa = 0.05$ and $\kappa = 0.0001$ are considered in order to rank the ES estimation models with respect to the traffic light system prescribed by the Basel Committee.

6.1. Financial data - S&P 500 index

For the analysis conducted within this chapter, daily closing prices of the S&P 500 index are taken from the *Thomson Reuters Datastream* database for the time period from 01.01.2006 to 01.10.2019. In order to remain consistent with the previous notation, the S&P 500 price data is transformed to a time series of log return losses. More precisely, if P_t and P_{t-1} denote the prices of the S&P 500 at times t and $t-1$, the respective log return loss is defined as,

$$l_t := -(\ln(P_t) - \ln(P_{t-1})). \tag{6.1}$$

Public holidays without any trading activities, i.e. with a related log return loss of exactly zero, are removed from the overall sample, as they would potentially bias the obtained backtesting results.
As shortly mentioned above, the overall sample is divided into two distinct time periods. The first one is labelled as the *crisis period* and contains a backtesting horizon of exactly 500 trading days, which spans from 27.12.2007 to 21.12.2009. All ES estimates, $\widehat{ES}_{t,0.975}$, are forecasted based on the rolling window of the previous 250 log return losses.[25] The crisis period includes the peak of the financial crisis in autumn 2008. Thus, this highly volatile market regime induces a major challenge for any of the applied risk estimation models.
The second considered backtesting period within this chapter contains the most recent S&P 500 data available at the start of the processing period for this thesis. Respectively, the period is labelled as *recent period* for the remainder of this thesis and includes 500 observations from 27.09.2017 to 23.09.2019. Again, all ES forecasts are estimated based on the rolling window of the previous 250 log return losses. [26] Compared to the crisis

[25] Given the rolling estimation window of 250 trading days, S&P 500 market data is considered starting from 28.12.2006, prior to the beginning of the crisis period, for estimation purposes.

[26] For the recent period, S&P 500 market data used for estimation purposes dates back until 28.09.2016.

period, the recent period displays a time frame with rather calm market conditions.
The differences in the volatility of the S&P 500 between the two considered backtesting periods can be observed in Figure 4, which depicts the log return losses over the respective time frames. Additionally, Figure 4 reveals that log return losses of the S&P 500 exhibit heteroskedasticity, which is a well known pattern of financial time series. Moreover, volatility clusters can be observed within both considered periods.
Another feature of log-return series is that they are often leptokurtic, i.e. they possess over-kurtosis compared to the standard normal distribution. In order to assess if log-return losses of the S&P 500 display leptokurtic patterns, Figure 5 depicts the QQ-plots of log return losses of the S&P 500 in percent against the standard normal distribution for both considered time periods. Especially in the crisis period, it is obvious that the considered log return losses posses fatter tails compared to the standard normal distribution, which can be observed on the left subplot of Figure 5. Even in the recent period, in a rather calm market environment, log return losses are still slightly leptokurtic. Nevertheless, distribution tails are by far thinner in the recent period compared to the crisis period.
Additionally, the differences in the distribution tails for both time windows of log return losses can be observed in Figure 6, which shows the respective histograms of log return losses. There are two further aspects which can be gathered from Figure 6. First of all, log return losses in both subplots of Figure 6 are centred around zero, which will be relevant for the subsection thereafter. Indeed, several of the applied ES estimation models are based on the assumption that average daily log return losses do not differ from zero. Secondly, log return losses over both considered backtesting time frames are slightly skewed to the right. As mentioned before, this is also a typical pattern of return loss series.
As outlined within this subsection, log return losses of the S&P 500 over both considered time frames exhibit typical patterns of financial time series. Overall, I decided to evaluate the ES backtests in scope on two distinct backtesting periods, as they display different market environments. On the one hand, in the recent period most ES estimation models are expected to work properly. Correspondingly, most ES estimation models should also pass any ES backtest over the recent period. On the other hand, it is far more challenging to make accurate ES forecasts in times of financial distress. Therefore, a majority of the ES estimation models might fail on the introduced backtests within the crisis period. Overall, regulators need to assure sufficient capital buffers irrespective of the underlying market environment. In a crisis period, a reasonable ES backtest needs to be penalizing and reject risk estimation models which do not quickly adopt to financial distress. On the contrary, in calm financial markets ES backtests should reward a sufficient risk coverage and not force a financial institution to hold unnecessarily high safety buffers.
Furthermore, the length of both considered backtesting horizons, i.e. exactly 500 observations, is motivated by the simulation results obtained within the previous chapter. As previously noticed, a backtesting horizon of 250 observations might be too short in order to consistently detect the risk underestimation inherited in an ES estimation model. On the other hand, excessively large backtesting windows might be problematic in practical application as they require the storage and moreover the availability of risk forecasts and return data over a long time frame. Thus, I believe the chosen backtesting horizon of $T = 500$ days gives a good trade off between backtesting power and data requirements, such that the conducted analysis displays a realistic set-up for an application by financial institutions or regulators.
The following subsection is first going to shortly summarize all five considered ES estimation models before the actual backtesting results can be discussed.

6.2. ES estimation models

This subsection presents all five applied ES estimation models, which are fitted to the S&P 500 log return loss data. All implemented ES estimation models differ with respect to their complexity and thus also yield different accuracies in modelling the underlying risk. Therefore, it is also interesting to observe, whether the applied ES backtests are able to discriminate between the different types of estimation models, in the following subsection.
The first, rather naive, approach corresponds to the ES estimation procedure applied within the size and power simulation studies. Thus, ES forecasts are based on a normal distribution which is fitted to the previous 250 observations. The second, third and fourth approach are also parametric ES estimation models, which apply econometric techniques in order to account for time varying conditional volatility of log return losses. Moreover, the respective econometric models all belong to the class of *Generalized Autoregressive Conditional Heteroskedasticity (GARCH)* processes, which were introduced by Bollerslev (1986). The three respective estimation models differ in their ability to capture the fat tails and the right-skewness of log return losses. The fifth ES estimation model is a *Filtered Historical Simulation (FHS)* with volatility updating, whereas conditional volatility is modelled with the RiskMetrics methodology suggested by JP Morgan (1996). Compared to all other considered models, the applied FHS it the only non-parametric approach and therefore requires no distributional assumption. As a consequence, comparable FHS methodologies are often applied by practitioners. In the following, the estimation procedure for all five approaches is shortly outlined.

(I) **ES-norm**
This first ES estimation model is labelled as *ES-norm*. Similar to the previous chapter, a normal distribution is fitted to the rolling window of the previous 250 S&P 500 log-return losses, such that $\widehat{L}_t \sim N(\widehat{\mu}_t, \widehat{\sigma}_t)$ for any observation t within the respective backtesting horizon. Moreover, the corresponding VaR and ES estimates can easily

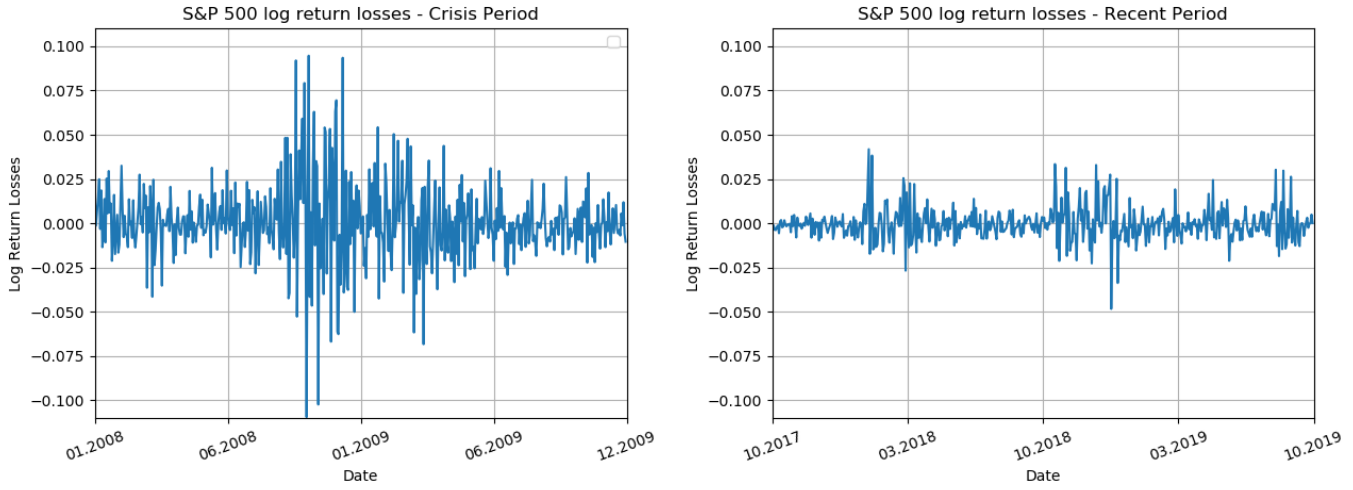


Figure 4: Plot of S&P 500 log return losses. Left subplot depicts the crisis period between 27.12.2007 and 21.12.2009, whereas the right subplot depicts the recent period between 27.09.2017 and 23.09.2019.

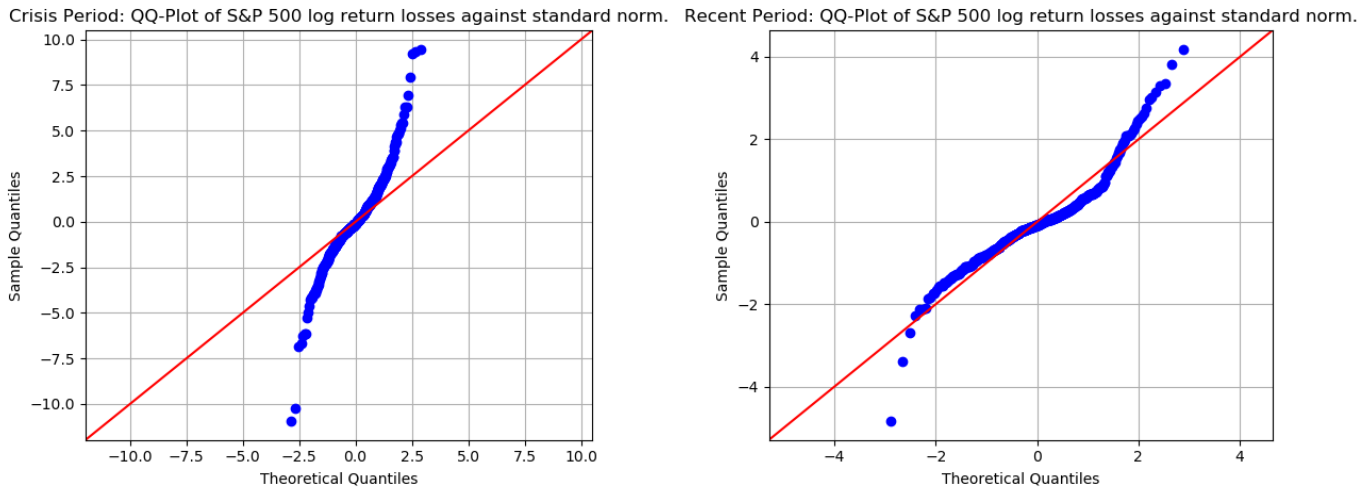


Figure 5: QQ-Plot of S&P 500 log return losses in percent against the standard normal distribution. Left subplot depicts the crisis period between 27.12.2007 and 21.12.2009, whereas the right subplot depicts the recent period between 27.09.2017 and 23.09.2019.

be calculated as,[27]

$$\widehat{VaR}_{t,\alpha} = \widehat{\mu}_t + \widehat{\sigma}_t \Phi^{-1}(\alpha), \tag{6.2}$$

$$\widehat{ES}_{t,\alpha} = \widehat{\mu}_t + \widehat{\sigma}_t \frac{\phi(\Phi^{-1}(\alpha))}{1-\alpha}, \tag{6.3}$$

where again ϕ and Φ denote the PDF and the CDF of the standard normal distribution, respectively. The ES-norm estimation model is a popular approach because of its simplicity. Nevertheless, risk estimates in the ES-norm model only slowly adopt to any changes in the volatility environment. Furthermore, especially in the evaluated crisis period, the distribution tails of S&P 500 log return losses heavily deviate from those of a standard normal distribution, which questions the assumption of normally distributed log return losses used in ES-norm. The resulting ES estimates and the respective S&P 500 log return losses, for both the crisis and the recent period, are depicted in Figure 7.

(II) **ES-GARCH-norm**

The second ES estimation model is labelled as *ES-GARCH-norm*. Compared to the first approach, conditional volatility is fitted to a GARCH(1,1) model in order to better capture the dynamics of volatility. For the second, as well as the third and fourth approach I assume that the daily conditional mean of log return losses

[27] See McNeil, Frey, and Embrechts (2015) pages 65 and 70 for a derivation of both formulas.

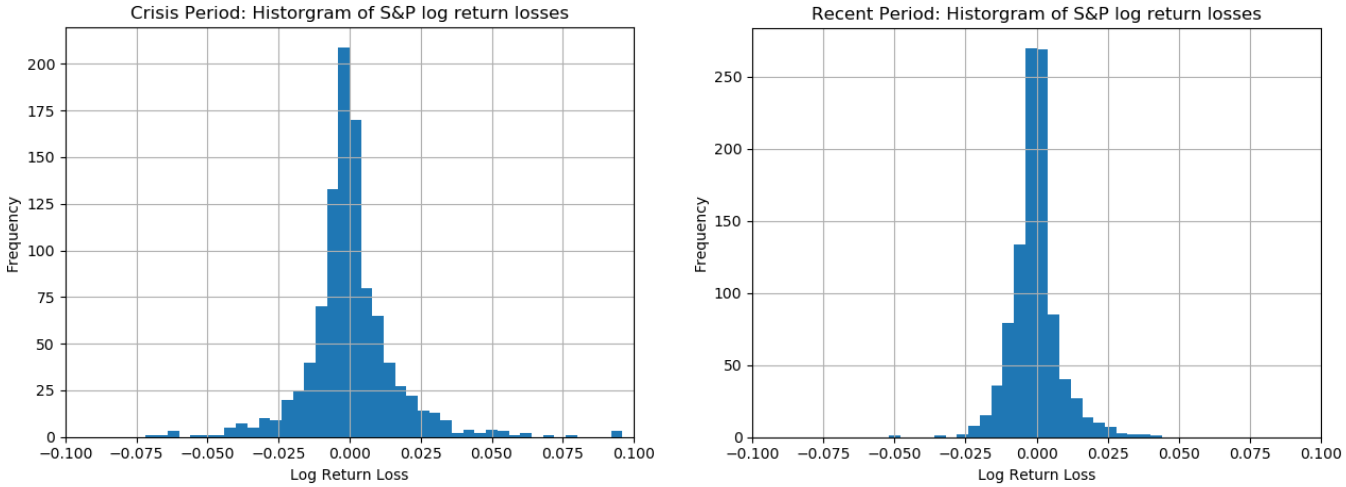


Figure 6: Histogram of S&P 500 log return losses. Left subplot depicts the crisis period between 27.12.2007 and 21.12.2009, whereas the right subplot depicts the recent period between 27.09.2017 and 23.09.2019.

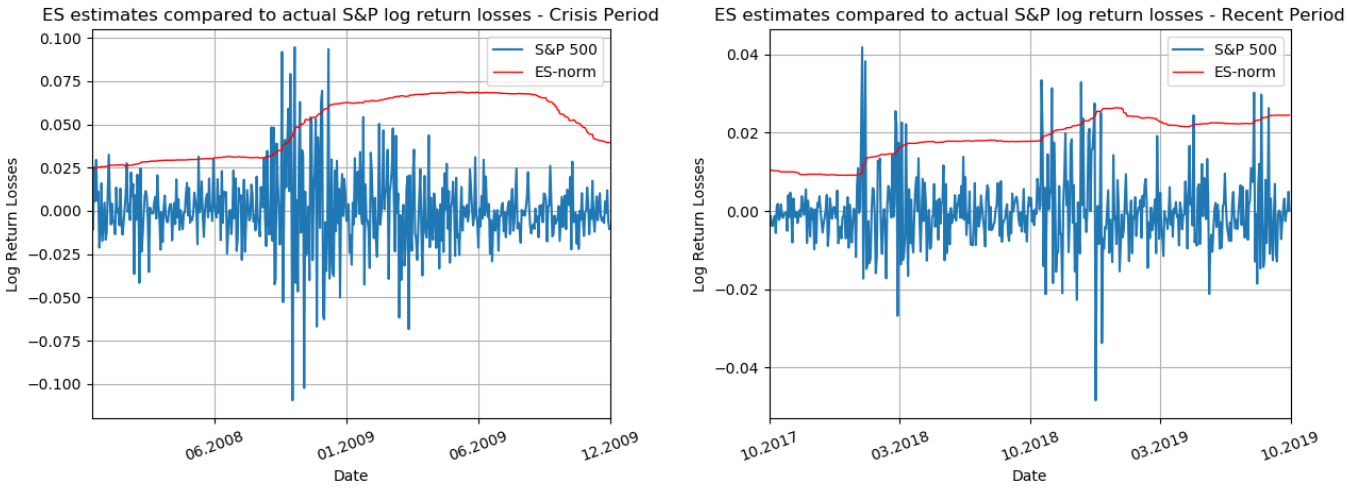


Figure 7: ES estimations based on fitted normal and t-distribution. Left subplot depicts the crisis period between 01.01.2007 and 01.01.2009, whereas the right subplot depicts the recent period between 01.10.2017 and 01.10.2019.

is zero, which is a reasonable assumption given the argumentation in the previous subsection. Moreover, log return losses at time t are specified by,

$$l_t = \sigma_t \varepsilon_t, \tag{6.4}$$

where σ_t again denotes the conditional volatility of log-return losses and ε_t is an innovation term. For the considered GARCH(1,1) model, innovations are assumed to be i.i.d. normally distributed, i.e. $\varepsilon_t \overset{i.i.d}{\sim} N(0,1)$. Furthermore, the volatility process is given by,

$$\sigma_t^2 = \omega + \alpha l_{t-1}^2 + \beta \sigma_{t-1}^2. \tag{6.5}$$

Thus, the volatility at any time t within the backtesting horizon depends on both the squared log return loss and the squared volatility from the previous time step.[28] For every observation t within the backtesting horizon, the parameter vector (ω, α, β) is fitted to the rolling window of the previous 250 log return losses, which is done in *Python* by applying a Maximum Likelihood (ML) estimation. Consecutively, one step ahead volatility forecasts $\widehat{\sigma}_t$ are calculated from the obtained parameters based on the volatility specification in (6.5). As outlined in McNeil et al. (2015) page 133, the respective VaR and ES

[28] As outlined above, the theory on a GARCH(1,1) model with normally distributed innovations ε_t was originally derived by Bollerslev (1986).

forecasts can be calculated as,

$$\widehat{VaR}_{t,\alpha} = \widehat{\sigma}_t q_\alpha(\varepsilon_t) = \widehat{\sigma}_t VaR_{t,\alpha}(\varepsilon_t), \quad (6.6)$$

$$\widehat{ES}_{t,\alpha} = \widehat{\sigma} ES_{t,\alpha}(\varepsilon_t). \quad (6.7)$$

Given the standard normal distribution of innovations, $VaR_{t,\alpha}(\varepsilon_t)$ and $ES_{t,\alpha}(\varepsilon_t)$ can be calculated according to formulas (6.2) and (6.3).
Overall, ES estimations obtained by the ES-GARCH-norm model are far more flexible to incorporate the current market environment compared to the previous rather static ES-norm approach. The resulting ES forecasts are depicted in Figure 8 for both the crisis and the recent period, together with the forecasts of all applied GARCH-type estimation models.

(III) **ES-GARCH-t**
The third ES estimation model, which is labelled as *ES-GARCH-t*, is closely related to the previous estimation approach. Again, log return losses are modelled by equation (6.4) and the conditional volatility is specified as in equation (6.5). In comparison to the previous ES estimation model, innovations ε_t are assumed to follow an i.i.d. standard t-distribution with ν degrees of freedom, i.e. $\varepsilon_t \overset{i.i.d.}{\sim} t_\nu$.
In this case not only the GARCH(1,1) parameters need to be estimated, but additionally also the degrees of freedom of the innovation term. Thus, for every t in the backtesting horizon the vector $(\nu, \omega, \alpha, \beta)$ is estimated in a ML procedure based on the rolling window of the previous 250 log return loss observations. Again, one step ahead volatility forecasts $\widehat{\sigma}_t$ are calculated based on the specified volatility process and the obtained GARCH(1,1) parameters. In the following, the VaR and ES forecasts can again be calculated according to formulas (6.6) and (6.7). In this case, it needs to be noted that innovations follow a standard t-distribution with ν degrees of freedom, such that the overall ES and VaR estimates in the ES-GARCH-t model are given by,

$$\widehat{VaR}_{t,\alpha} = \widehat{\sigma}_t VaR_{t,\alpha}(\varepsilon_t) = \widehat{\sigma}_t t_\nu^{-1}(\alpha), \quad (6.8)$$

$$\widehat{ES}_{t,\alpha} = \widehat{\sigma}_t ES_{t,\alpha}(\varepsilon_t) = \widehat{\sigma}_t \frac{g_\nu(t_\nu^{-1}(\alpha))}{1-\alpha}\left(\frac{\nu + (t_\nu^{-1}(\alpha))^2}{\nu - 1}\right), \quad (6.9)$$

where g_ν and t_ν denote the PDF and the CDF of the standard t-distribution, respectively.[29]
The resulting ES forecasts for the ES-GARCH-t model are also depicted in Figure 8, for both the crisis and the recent period. From a theoretical point of view, the ES-GARCH-t approach is superior in accounting for the fat tails of the S&P 500 log return losses compared to the ES-GARCH-norm model. More generally speaking, Angelidis, Benos, and Degiannakis (2004) for example concludes, that GARCH models with heavy tailed innovation distributions should be preferred as they better capture the leptokurtic behaviour of financial data.

(IV) **ES-EGARCH-t**
The fourth considered ES estimation model is labelled as *ES-EGARCH-t*. Compared to the previous two models, the ES-EGARCH-t uses an extension of the classical GARCH framework in order to model conditional volatility. Moreover an *exponential GARCH (EGARCH)* is applied, which was originally introduced by Nelson (1991). For the selected EGARCH(1,1) model specification, log return losses are again specified as in formula (6.4), whereas innovations are assumed to follow an i.i.d. standard t-distribution with ν degrees of freedom, i.e. $\varepsilon_t \overset{i.i.d}{\sim} t_\nu$. In comparison to the previous two models, the conditional volatility process is defined through,

$$h_t = \omega + \alpha \frac{l_{t-1}}{\sqrt{\sigma_{t-1}^2}} + \gamma\left(\left|\frac{l_{t-1}}{\sqrt{\sigma_{t-1}^2}}\right| - \mathbb{E}\left[\left|\frac{l_{t-1}}{\sqrt{\sigma_{t-1}^2}}\right|\right]\right) + \beta h_{t-1}, \quad (6.10)$$

where $h_t = \ln(\sigma_t^2)$. Similar as for the previous models, all relevant parameters of the EGARCH specification are fitted for every t within the backtesting horizon on the rolling window of the preceding 250 log return losses. In this case the parameter vector $(\nu, \omega, \alpha, \gamma, \beta)$ is fitted in every ML estimation. Afterwards, one step ahead volatility forecasts $\widehat{\sigma}_t$ based on equation (6.10) and the obtained parameter values are calculated. Moreover, the respective VaR and ES forecasts for the ES-EGARCH-t model are obtained analogous to equations (6.8) and (6.9), as innovations are assumed to follow an i.i.d standard t-distribution with ν degrees of freedom.
From a theoretical perspective, the EGARCH model is able to capture the so called *leverage effect*, which is often observed in financial time series and which is the main cause for right-skewed log returns loss series. In practice, volatility often increases more following a high return loss compared to a gain of the same magnitude. In the GARCH methodology the past log return losses have a symmetric impact on conditional volatility, which can be recognized from equation (6.5). The EGARCH methodology on the contrary, allows to model an asymmetric impact depending on the sign of observed log return losses. Amongst researchers, the EGARCH(1,1) specifications is widely used in various applications to capture the behaviour of conditional volatility.
The ES estimates of the ES-EGARCH-t model, for both the crisis and the recent period, are also depicted in Figure 8.

[29] See McNeil et al. (2015) pages 66 and 71 for a detailed description of the calculation of $VaR_{t,\alpha}(\varepsilon_t)$ and $ES_{t,\alpha}(\varepsilon_t)$ given that $\varepsilon_t \overset{i.i.d}{\sim} t_\nu$.

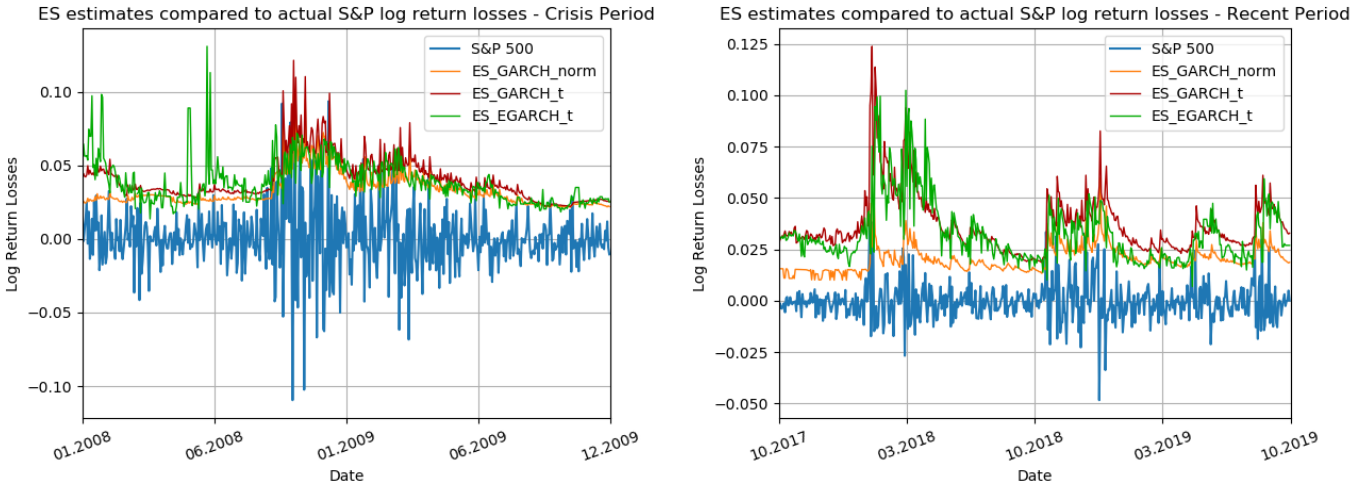


Figure 8: ES estimations based on GARCH-type specifications of conditional volatility. Left subplot depicts the crisis period between 01.01.2007 and 01.01.2009, whereas the right subplot depicts the recent period between 01.10.2017 and 01.10.2019.

Overall, all three GARCH-type ES forecasts show similar patterns over time. Indeed, both approaches based on t-distributed innovations show slightly more conservative risk forecasts compared to the ES-GARCH-norm model, which might be due to a superior ability to capture the fat tails of log return losses. This becomes perceptible as risk estimates of the ES-GARCH-norm model are in general on a lower level compared to the other two approaches. Furthermore, there are barely any differences in the forecasts of both the ES-GARCH-t and the ES-EGARCH-t model, which might be caused by the rather low skewness of the underlying S&P 500 data. Indeed, during the peak of the crisis period, the ES-GARCH-t model might even yield a slightly superior coverage of the underlying risk.

(V) **ES-FHS-RiskMetrics**

The fifth ES estimation model is labelled as *ES-FHS-RiskMetrics*. Compared to the previous four models, it yields a non-parametric approach and is thus especially popular amongst practitioners. It makes use of a *Filtered Historical Simulation (FHS)* based on volatility forecasts given by the RiskMetrics approach outlined in JP Morgan (1996). Moreover, in the RiskMetrics methodology volatility is estimated by the use of an exponentially weighted moving average (EWMA) scheme. For every observation t within the backtesting horizon, the respective volatility is given by,

$$\sigma_t^2 = \lambda\sigma_{t-1}^2 + (1-\lambda)l_{t-1}^2, \tag{6.11}$$

where λ is the decay factor, which calibrates the weights in the EWMA scheme. In line with the recommendation given by JP Morgan (1996), I set the decay factor to a value of $\lambda = 0.94$. For every observation t within the backtesting horizon, again the previous 250 return loss observations are used for estimation purposes. In detail, one can calculated the volatility of all 250 preceding observations, i.e. $(\sigma_{t-250},\ldots,\sigma_{t-1})$, as well as a forecast for the volatility at time t, i.e. $\widehat{\sigma}_t$, based on equation (6.11). For every time t within the backtesting horizon, the sample $S_{\text{FHS},t}$ functions as a basis for the FHS and is defined as,

$$\begin{aligned} S_{\text{FHS},t} &:= \left\{ l_s \frac{\widehat{\sigma}_t}{\sigma_s}, \text{ for } s = t-250,\ldots,t-1 \right\} \\ &:= \{ l_s^{\text{scaled}}, \text{ for } s = t-250,\ldots,t-1 \}. \end{aligned} \tag{6.12}$$

Thus, $S_{\text{FHS},t}$ contains all log return losses over the preceding estimation period, which are scaled according to changes in the conditional volatility over time. If volatility was lower at time $s < t$ compared to the value forecasted for time t, the respective log return loss is up-scaled in order to capture the increased riskiness inherited in the financial market. The same also holds true vice-versa if volatility decreases over time. In order to calculate both VaR and ES forecasts in the ES-FHS-RiskMetrics approach, empirical quantiles of $S_{\text{FHS},t}$ are exploited in the following way,

$$\widehat{VaR}_{t,\alpha} = \tilde{q}_\alpha(S_{\text{FHS},t}), \tag{6.13}$$

$$\widehat{ES}_{t,\alpha} = \frac{1}{\sum_{s=1}^{250} \mathbb{1}_{\{l_{t-s}^{\text{scaled}} > \tilde{q}_\alpha(S_{\text{FHS},t})\}}} \sum_{t=1}^{250} l_{t-s}^{\text{scaled}} \mathbb{1}_{\{l_{t-s}^{\text{scaled}} > \tilde{q}_\alpha(S_{\text{FHS},t})\}}, \tag{6.14}$$

where $\tilde{q}_\alpha(S_{\text{FHS},t})$ denotes the empirical α-quantile of $S_{\text{FHS},t}$.[30] The resulting ES estimates of the ES-FHS-RiskMetrics approach for both considered periods of

[30] More details on the methodology of a FHS can for example be found in Hull (2015), chapter 13, or in McNeil et al. (2015), chapter 9.2.

S&P 500 log return losses are depicted in Figure 9 below. Compared to the previously stated GARCH methodologies, the ES-FHS-RiskMetrics approach yields slightly less volatile risk estimates. Nevertheless, the magnitude of obtained ES forecasts still appears to be sufficient at a first glance, to cover the underlying risk in both considered periods.

As outlined within this subsection, all five introduced forecasting models differ with respect to their complexity and methodology and thus also yield a different quality of obtained ES estimates for the underlying series of S&P 500 market data. Nevertheless, apart from the rather naive ES-norm model, all other estimation approaches might potentially also be implemented within financial institutions in a similar manner. Therefore, the judgement on the obtained ES forecasts displays a realistic example of use for the ES backtests in scope.

6.3. Backtesting results

This subsection is going to evaluate all estimated ES figures for both considered time periods of S&P 500 log return losses based on all outlined ES backtests. Furthermore, in line with the main objective of this thesis, a one-sided test design will be applied in order to detect estimation models which underestimate the true risk inherited in the S&P 500 index.

For the purpose of this subsection, only the de-facto one-sided version of the multinomial testing framework will be applied given the selected one-sided backtesting design. This can be justified, as both multinomial versions displayed almost identical size and power figures in the previous chapter, while the one-sided version additionally rules out the rejection of any conservative ES estimation model. Furthermore, the approximate Z-test proposed within this thesis will be applied instead of the original approach outlined by Costanzino and Curran (2015), as the approximative version exhibited slightly superior power and size values in the preceding simulation chapter. In addition, the approximative version can also be applied to any of the outlined ES estimation models given that it does not require explicit parametric estimates of the distribution of return losses, as the original version does. All other backtesting approaches, i.e. the Test 2 from Acerbi and Szekely (2014), the intercept ESR backtest motivated by Bayer and Dimitriadis (2019) and the combined ES residuals backtest based on McNeil and Frey (2000), are implemented as described in chapter 4 of this thesis.

For both backtesting periods of the S&P 500 index, all five estimated ES models are backtested at significance levels of $\kappa = 0.05$ and $\kappa = 0.0001$. Both significance levels determine the traffic light system prescribed by the Basel Committee for the VaR traffic light test, as outlined in section 3.2. In the same fashion, I also want to group each of the five ES estimation models into the same colour scheme with respect to a certain ES backtest. Thus, if an estimation model is rejected at both considered significance levels it is assigned a red traffic light. An ES estimation model is labelled in yellow if it is rejected at $\kappa = 0.05$, but passes the respective ES backtest at a significance level of $\kappa = 0.0001$. An ES estimation model which passes the respective ES backtest at both considered significance levels is labelled in green.

In the following, subsection 6.3.1 presents the backtesting results over the crisis period, while subsection 6.3.2 thereafter summarizes the respective results over the recent period.

6.3.1. Crisis period

As outlined before, the crisis period contains the peak of the financial crisis in 2008, and thus exhibits an environment of financial distress which imposes a major challenge to any ES estimation model. An overview, of the performance of all five ES estimation models is depicted in Figure 10, which shows scatter plots of S&P 500 log return losses together with both estimated ES and VaR figures for each of the considered ES estimation models. Moreover, a log return loss l_t is marked in orange if it exceeds the respective $\widehat{VaR}_{t,0.975}$ threshold and it is marked in red if it additionally also violates the respective $\widehat{ES}_{t,0.975}$ estimate.

For all five estimation models, the majority of violations of both the VaR and the ES can be observed between June 2008 and January 2009. Indeed, the S&P 500 suffered the largest losses in the crisis period within this timespan, driven by the default of the Lehman Brothers at the 15.09.2008. The date when Lehman Brothers declared bankruptcy is market with dotted black lines in Figure 10. At a first glance, the ES-norm estimation model shows the worst results in this timespan with respect to both the number and the magnitude of observed outliers. This is not surprising, as both VaR and ES forecasts in the ES-norm model only slowly react to changes in the market environment. Around this peak of the financial crisis, all three GARCH-typ models exhibit similar patterns, whereas the ES-EGARCH-t model apparently displays slightly more severe violations compared to the other two approaches. From a first impression, the non-parametric ES-FHS-RiskMetrics approach yields the lowest degree of risk underestimation during the crisis period, with few violations, which additionally also appear to be of rather small magnitude.

In order to achieve a better understanding of the quality of the obtained risk estimates, I want to analyse some further statistics before turning to the actual ES backtesting results. As outlined within chapter 3, two aspects of ES estimates need to be taken into account in order to verify the correct unconditional coverage of an ES estimation model. First of all, the number of violations beyond $\widehat{VaR}_{t,0.975}$ over the backtesting horizon and secondly their magnitude with respect to the estimated $\widehat{ES}_{t,0.975}$. As outlined in Proposition 3.5, the first aspect is equivalent to the condition $\mathbb{E}[I_t(\alpha)] = 1-\alpha$ for every observation t within the backtesting horizon. Therefore, over the crisis period the expected number of VaR violation of an accurate estimation model is given by,

$$\begin{aligned}\mathbb{E}[\#\text{ of VaR violations}] &= \sum_{t=1}^{500} \mathbb{E}[I_t(\alpha)] \\ &= 500 \cdot 0.025 = 12.5. \quad (6.15)\end{aligned}$$

With respect to the second aspect listed above, the magnitude of outliers beyond the respective VaR threshold needs to be taken

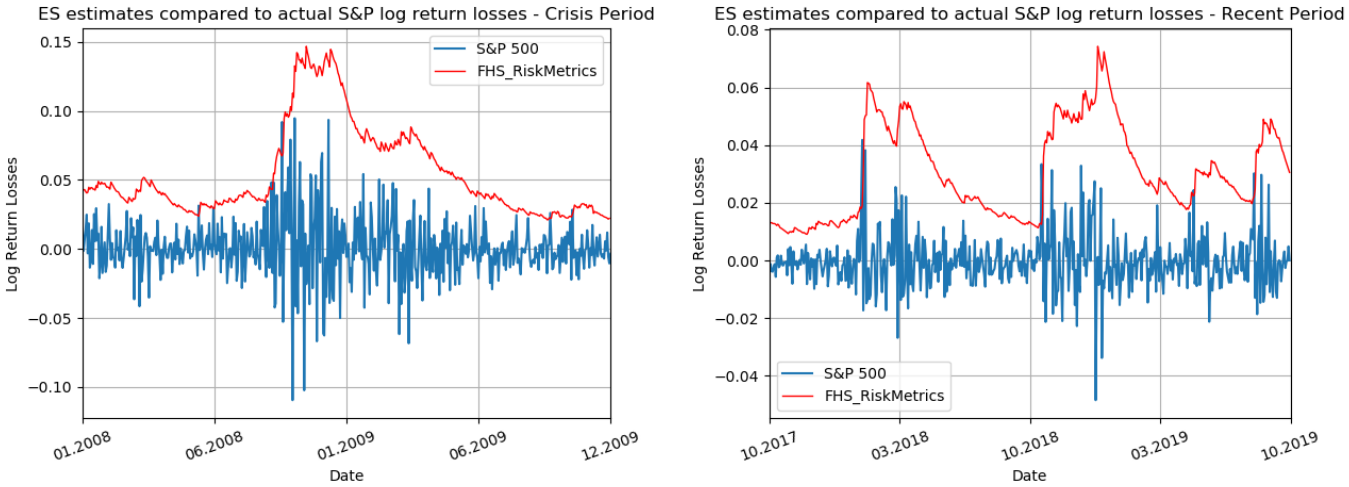


Figure 9: ES estimations - Filtered Historical Simulation with volatility updating based on RiskMetrics. Left subplot depicts the crisis period between 01.01.2007 and 01.01.2009, whereas the right subplot depicts the recent period between 01.10.2017 and 01.10.2019.

Crisis period - Violations of $\widehat{VaR}_{t,0.975}$ and $\widehat{ES}_{t,0.975}$

ES estimation model	# of VaR violations	# of ES violations	Average exceedance of ES in %
ES-norm	28	21	+24.44
ES-GARCH-norm	47	30	+18.19
ES-GARCH-t	32	6	+8.57
ES-EGARCH-t	34	14	+13.52
ES-FHS-RiskMetrics	10	6	+7.14

Table 13: Observed number of violations beyond $\widehat{VaR}_{t,0.975}$ and beyond $\widehat{ES}_{t,0.975}$, for all considered estimation models within the crisis period of the S&P 500. Furthermore, the average exceedance of VaR violations beyond the ES in percent is depicted.

into account. For an accurate ES estimation model, the average value of all VaR violations should be fairly close to the estimated ES figure. If on average VaR violations exceed the respective ES forecast, this hints towards a potential risk underestimation. For a first assessment of this second aspect, I define the *average exceedance of the ES* as,

$$\frac{1}{\sum_{t=1}^{500} I_t(\alpha)} \sum_{t=1}^{500} \frac{l_t - \widehat{ES}_{t,\alpha}}{\widehat{ES}_{t,\alpha}} I_t(\alpha) \tag{6.16}$$

In the following, both the observed number of VaR and ES violations as well as the average exceedance of the ES are listed in Table 13, for all ES estimation models within the crisis period. As depicted in Table 13, only the ES-FHS-RiskMetrics model exhibits less VaR violations than expected over the crisis period. All other applied estimation approaches fail to correctly model the conditional 0.975-quantile of log return losses during the period of high financial distress. Indeed, all four other estimation models display more than twice as many VaR outliers as expected. As previously noted, the magnitude of violations is the largest for the ES-norm model where 21 out of 28 VaR violations also lie beyond the estimated ES forecast, which leads to an average exceedance of the ES of 24.44 %. With respect to the magnitude of outliers, also all other ES estimation models underestimate the underlying risk during the crisis period. The least degree of risk underestimation in the distribution tail beyond the conditional 0.975-quantile can be observed for both the ES-FHS-RiskMetrics and the ES-GARCH-t model, with average exceedance values of the ES below 10 %. Based on this first assessment, I expect that the ES-FHS-RiskMetrics model achieves the best backtesting results in the crisis period. All other ES estimation approaches display weak results, especially in forecasting the conditional 0.975-quantile of log return losses given by $\widehat{VaR}_{t,0.975}$.

The actual backtesting results for all considered ES backtests over the crisis period of S&P 500 log return losses are listed in Table 14. Overall, the achieved backtesting results are in line with expectations given the preceding argumentation. Furthermore, there are no major differences regarding the traffic light classification of ES estimation models between the applied backtests. The ES-FHS-RiskMetrics approach passes all conducted backtests at both considered significance levels and is thus universally labelled in green over the crisis period. Nevertheless, achieved p-values for the ES-FHS-RiskMetrics model show a high variation between roughly 9% and 80%. Although

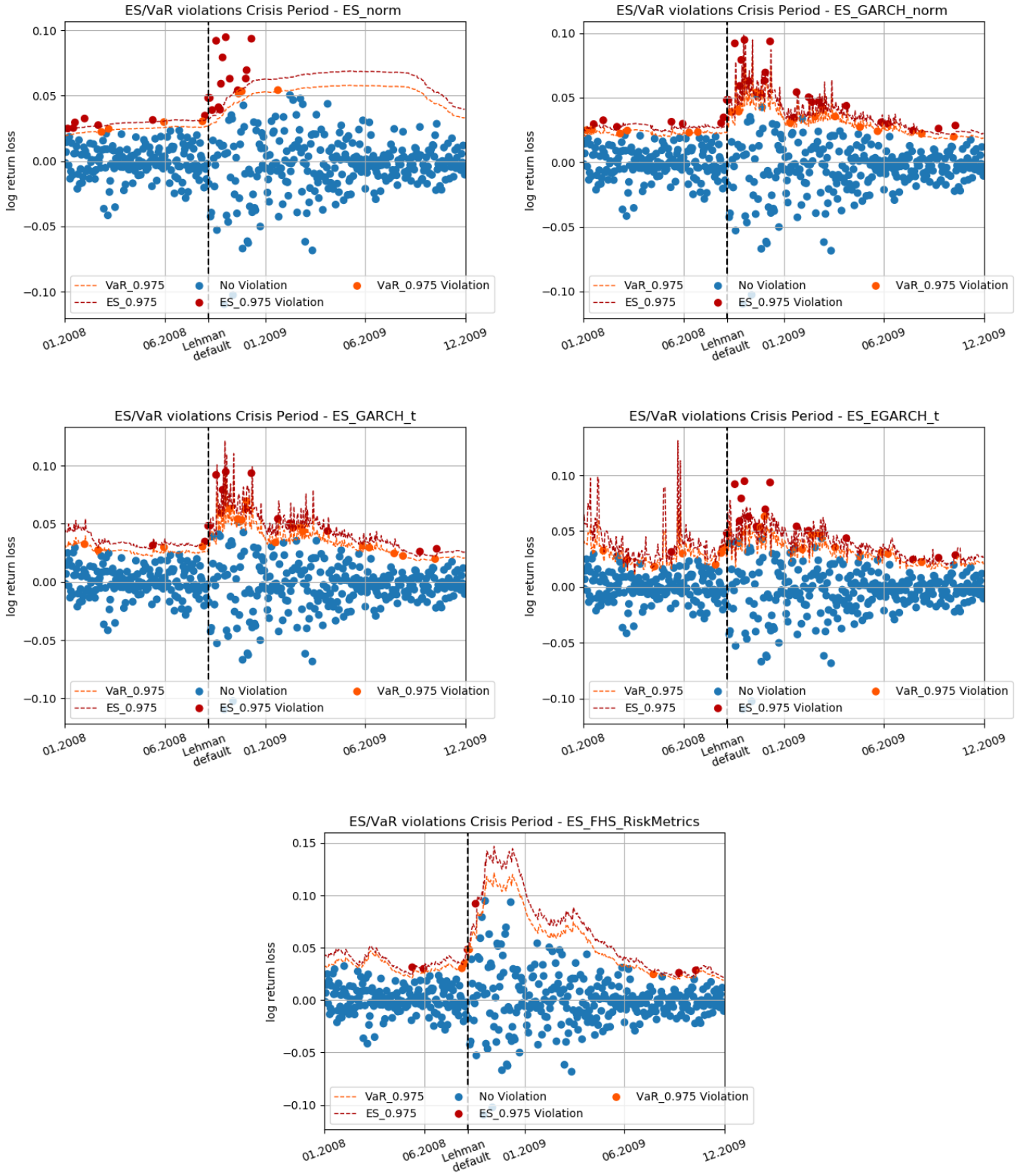


Figure 10: Scatter plots with observed log return losses from crisis period of the S&P 500. Violations beyond $\widehat{VaR}_{t,0.975}$ are marked in orange and violations beyond $\widehat{ES}_{t,0.975}$ in red for all five considered ES estimation models.

Crisis Period - ES backtesting results

		Estimation model				
ES backtest	Significance Level	ES-norm	ES-GARCH-norm	ES-GARCH-t	ES-EGARCH-t	ES-FHS-RiskMetrics
Multinomial one-sided	0.05	✗	✗	✗	✗	✓
	0.0001	✗	✗	✗	✗	✓
	p-value	< 0.0001	< 0.0001	< 0.0001	< 0.0001	0.0964
	traffic light	**red**	**red**	**red**	**red**	**green**
Test 2	0.05	✗	✗	✗	✗	✓
	0.0001	✗	✗	✗	✗	✓
	p-value*	< 0.0001	< 0.0001	< 0.0001	< 0.0001	0.804
	traffic light	**red**	**red**	**red**	**red**	**green**
Intercept ESR	0.05	✗	✗	✗	✗	✓
	0.0001	✗	✗	✓	✗	✓
	p-value	< 0.0001	< 0.0001	0.0100	< 0.0001	0.1300
	traffic light	**red**	**red**	**yellow**	**red**	**green**
Z-test approximative version	0.05	✗	✗	✗	✗	✓
	0.0001	✗	✗	✗	✗	✓
	p-value	< 0.0001	< 0.0001	< 0.0001	< 0.0001	0.3786
	traffic light	**red**	**red**	**red**	**red**	**green**
Combined ES residuals	0.05	✗	✗	✗	✗	✓
	0.0001	✗	✗	✗	✗	✓
	p-value	< 0.0001	< 0.0001	< 0.0001	< 0.0001	0.1160
	traffic light	**red**	**red**	**red**	**red**	**green**

Table 14: Backtesting results of all five implemented ES estimation models with respect to all ES backtests in scope over the crisis period of S&P log-return losses. Backtests are conducted at significance levels of $\kappa = 0.05$ and $\kappa = 0.0001$. Results are labelled with an ✗if the H0 in the one-sided test is rejected an with an ✓if the H0 can not be rejected.
*p-values for the Test 2 correspond to the bootstrap test version applied only for $\kappa = 0.0001$.

Recent period - Violations beyond $\widehat{VaR}_{t,0.975}$ and $\widehat{ES}_{t,0.975}$

ES estimation model	# of violations of $\widehat{VaR}_{t,0.975}$	# of violations of $\widehat{ES}_{t,0.975}$	Average exceedance of ES in %
ES-norm	28	19	+39.57
ES-GARCH-norm	19	11	+24.21
ES-GARCH-t	15	1	-4.98
ES-EGARCH-t	19	6	+11.26
ES-FHS-RiskMetrics	16	6	+7.20

Table 15: Observed number of violations beyond $\widehat{VaR}_{t,0.975}$ and beyond $\widehat{ES}_{t,0.975}$ as well as average exceedance of VaR violations beyond the ES in percent, for all considered estimation models within the recent period.

this has no impact on the assigned traffic light, test decisions might diverge if further backtesting significance levels are taken into account.

On the contrary, all other implemented ES estimation models are rejected at all applied backtests at a significance level of $\kappa = 0.05$, and at almost all applied backtest at the more extreme significance level of $\kappa = 0.0001$. Indeed, only the intercept ESR backtest does not reject the ES-GARCH-t estimation model at the more extreme significance level. Therefore, the ES-GARCH-t estimation model achieves four red and one yellow traffic light, while the three remaining ES estimation models are assigned a red label throughout all backtests in the crisis period. This minor difference in assigned traffic lights might be justified with the results listed in Table 13, as the ES-GARCH-t estimation model apparently displays a lower degree of risk underestimation compared to the ES-norm, the ES-GARCH-norm and the ES-EGARCH-t approache. From a theoretical perspective, it is also reasonable that the ES-GARCH-t approach outperforms both estimation models based on normality assumptions, due to its superior ability to capture the highly leptokurtic patter of S&P 500 log return losses in the crisis period.

It should be noted that for the rather extreme value of $\kappa = 0.0001$ results of all applied bootstrap test decisions might potentially change depending on the respective simulation output. For the more extreme significance threshold, the Test 2[31], the combined ES residuals backtest as well as the intercept ESR backtest are based on a simulation procedure. Given the extreme significance threshold, the output of single simulations trials might decide over the achieved test decision unless a very large amount of simulation trials M is considered. As results depicted in Table 14 are overall plausible, I stick with values of $M = 1000$ for both the Test 2 and the combined ES residuals backtest as well as $M = 100$ for the intercept ESR backtest following the argumentation given chapter 4. Furthermore, the selected number of simulation trials assures that computational times stay within an acceptable range for practical applications. Nevertheless, one should be aware of this issue, whenever a bootstrap decision is applied at rather extreme significance levels.

[31] As outlined within chapter 4.2 the Test 2 from Acerbi and Szekely (2014) is conducted with fixed critical values for a significance level of $\kappa = 0.05$, whereas a bootstrap decision is applied for a value of $\kappa = 0.0001$.

6.3.2. Recent period

This subsection presents the backtesting results over the most recent period of S&P 500 log return losses available at the start of the processing period for this thesis. As previously outlined the recent period depicts a rather calm market environment compared to the previously analysed crisis period. Thus, also the implemented ES estimation models are expected to show superior results over this second backtesting period.

A first impression of the accuracy of ES estimates can be gathered from Figure 11, which depicts the scatter plots of S&P 500 log return losses together with the respective VaR and ES estimates. Again all observed violations are marked in the same fashion as in Figure 10. Similar as in the crisis period, the ES-norm model is not able to capture the rather moderate peaks in log-return losses, due to its limited reactivity to changing market conditions. Again the patterns for all applied GARCH models look similar at a first glance. Nevertheless, apparently the ES-GARCH-t model yields the lowest amount of outliers amongst these three models. The non-parametric ES-FHS-RiskMetrics model again produces slightly less volatile ES estimates compared. Still, both the number and the magnitude of violations appears to be comparable to the implemented GARCH methodologies.

In order to get an even better idea of the performance of all implemented estimation models, Table 15 again lists some additional statistics regarding the number and the magnitude of observed violations. In the recent period, all estimation models exceed the expected number of VaR violations of 12.5 derived in formula (6.15). Nevertheless, four out of five estimation models range within a value of 15 to 19 observed VaR violations, which is still reasonably low. The ES-norm model on the contrary again realizes more than twice as many outliers as expected. Regarding the magnitude of VaR violations, both the ES-norm and the ES-GARCH-norm model stand out with an average exceedance of the ES of +39.57 % and +24.21 %. Both values indicate towards a rather high degree of risk underestimation. On the other hand, the ES-EGARCH-t and the ES-FHS-RiskMetrics approach only show a moderate degree of risk underestimation in the distribution tail with an average exceedance of the ES of +11.26 % and +7.20% respectively, whereas the ES-GARCH-t model is even slightly on the conservative side with a value of -4.98%.

Based on this first assessment, it is to be expected that the ES-

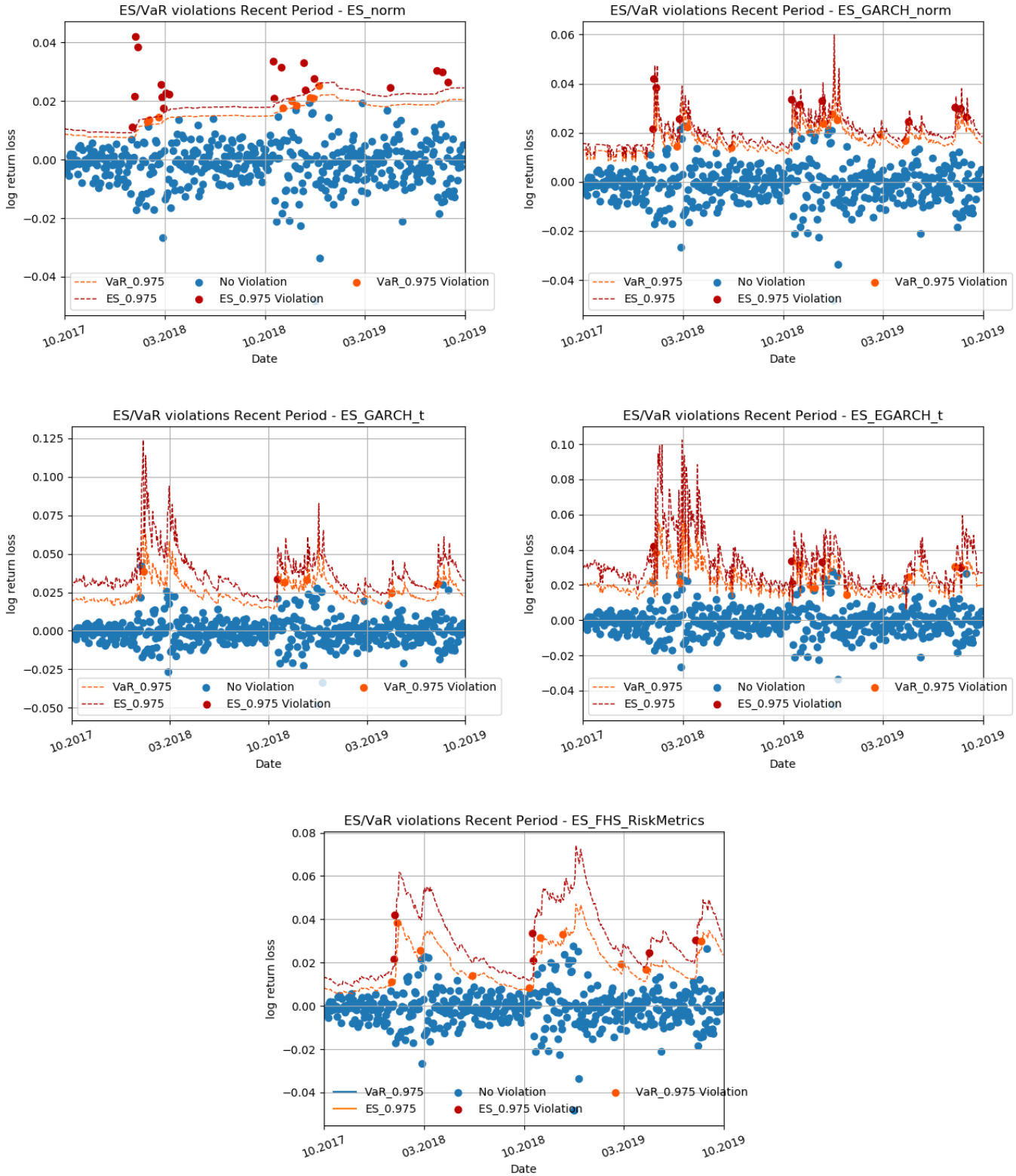


Figure 11: Scatter plots with observed log return losses from recent time period of the S&P 500. Violations beyond $\widehat{VaR}_{t,0.975}$ are marked in orange and violations beyond $\widehat{ES}_{t,0.975}$ in red for all five considered ES estimation models.

Recent Period - ES backtesting results

		Estimation model				
ES backtest	Significance Level	ES-norm	ES-GARCH-norm	ES-GARCH-t	ES-EGARCH-t	ES-FHS-RiskMetrics
Multinomial one-sided	0.05	✗	✗	✓	✓	✓
	0.0001	✗	✗	✓	✓	✓
	p-value	< 0.0001	< 0.0001	0.6416	0.5853	0.3683
	traffic light	**red**	**red**	**green**	**green**	**green**
Test 2	0.05	✗	✗	✓	✓	✓
	0.0001	✗	✗	✓	✓	✓
	p-value*	< 0.0001	< 0.0001	0.9840	0.1800	0.1970
	traffic light	**red**	**red**	**green**	**green**	**green**
Intercept ESR	0.05	✗	✗	✓	✗	✓
	0.0001	✗	✗	✓	✗	✓
	p-value	< 0.0001	< 0.0001	0.6700	< 0.0001	0.0700
	traffic light	**red**	**red**	**green**	**red**	**green**
Z-test approximative version	0.05	✗	✗	✓	✓	✗
	0.0001	✗	✓	✓	✓	✓
	p-value	< 0.0001	0.0003	0.9311	0.0775	0.0365
	traffic light	**red**	**yellow**	**green**	**green**	**yellow**
Combined ES residuals	0.05	✗	✗	✓	✓	✓
	0.0001	✗	✓	✓	✓	✓
	p-value	< 0.0001	0.0110	0.8240	0.1910	0.1279
	traffic light	**red**	**yellow**	**green**	**green**	**green**

Table 16:
Backtesting results of all five implemented ES estimation models with respect to all ES backtests in scope over the recent period of S&P log-return losses. Backtests are conducted at significance levels of $\kappa = 0.05$ and $\kappa = 0.0001$. Results are labelled with an ✗if the H0 in the one-sided test is rejected an with an ✓if the H0 can not be rejected.
*p-values for the Test 2 correspond to the bootstrap test version applied only for $\kappa = 0.0001$.

GARCH-t model performs best in the conducted backtests over the recent period followed by both the ES-FHS-RiskMetrics and the ES-EGARCH-t model. It is interesting to observe, that the ES-GARCH-t model apparently yields slightly superior ES estimates compared to the more sophisticated ES-EGARCH-t approach. This might be explained as in the recent period of S&P 500 data, only a weak form of the leverage effect can be observed, as log return losses are only slightly right-skewed. Thus, the increased estimation uncertainty in the ES-EGARCH-t model, due to a higher number of estimation parameters, might outweigh the theoretical advantage to capture a potential leverage effect. The two remaining estimation approaches, i.e. the ES-norm and the ES-GARCH-norm model, again fail to consistently forecast the underlying risk in the S&P 500 index. Whereas the former one is not able to react to changing market conditions, the later one mainly fails due to its inability to account for the fat tails of the S&P 500 market data.

The results of the five implemented models with respect to the ES backtests in scope over the recent period of S&P 500 log return losses are listed in Table 16 below. Overall, the backtesting results over the recent period are again in line with expectations. Both the ES-norm as well as the ES-GARCH-norm model are rejected in the majority of all applied ES backtests. Whereas the former receives five red traffic lights, the latter ends up with three red and two yellow classifications. This is in line with expectations, as the ES-GARCH-norm model performs slightly better compared to the ES-norm approach, based on the results depicted in Table 15.

On the contrary, all three remaining estimation models are labelled in green for most of the considered backtesting methodologies. In line with previous argumentation, the ES-GARCH-t approach yields the best backtesting results over the recent period and achieves the optimum of five green traffic lights. Moreover, all p-values related to the ES-GARCH-t model are very high with values above 64 %. Furthermore, the decent backtesting results of both the ES-EGARCH-t and the ES-FHS-RiskMetrics are also reasonable, given the rather moderate degree of risk underestimation depicted in Table 15. The only slightly surprising result is that the intercept ESR backtests rejects the ES-EGARCH-t approach over the recent period not just at a significance level of $\kappa = 0.05$ but also at the more extreme threshold of $\kappa = 0.0001$. Nevertheless, the majority of all test decisions in the recent period coincides with the considerations based on the number and magnitude of violations.

Concluding this chapter, the outlined backtesting set-up displays a realistic example for a real-world application. The applied backtesting period of $T = 500$ observations is a reasonable choice, with respect to the length of the required input time series. Furthermore, ES backtest versions are applied which yield a one-sided test decision compliant with regulatory needs and which are applicable for all common ES estimation models. Overall, all five implemented ES backtests reveal reasonable results over both considered time periods of S&P 500 data. Moreover, the achieved backtesting results are in line with expectations taking into account previous considerations regarding the underlying market conditions. Besides the naive ES-norm estimation model, all other backtested forecasts stem from realistic estimation models, which are indeed relevant for practical applications. Furthermore, all applied backtests are able to differentiate between varying degrees of modelling accuracy induced by the different estimation models fitted to the S&P 500 data. Overall, it is also straightforward to implement a traffic light system, stipulated by the Basel Committee, for all evaluated ES backtests. Furthermore, in the majority of all cases the traffic light assigned to a certain estimation model coincides across all conducted backtests. Although it should be noted, that p-values might substantially differ depending on the applied backtest, which might be relevant if a more precise classification of estimation models needs to be conducted compared to the traffic light approach. Furthermore, any bootstrap decisions, especially at extreme significance levels like $\kappa = 0.0001$, up to a certain extend depends on a random factor based on the nature of the simulation procedure. Thus, at rather extreme significance levels bootstrap decisions are subject to potential changes depending on the simulation output. Apart from the last two mentioned aspects, there are indeed no major issues which might hinder the implementation of the outlined ES backtesting versions in real world scenarios.

7. Conclusion

This thesis revealed that backtesting the ES is indeed not much more complicated than backtesting the VaR, despite all doubts and reservations expressed within the literature. Definitely, backtesting the ES somehow requires a higher degree of creativity and there will probably never be any conceptual fully untainted backtest for the ES compared to available approaches to backtest the VaR. Nevertheless, there are several promising backtesting approaches, which have been proposed within the last two decades. Overall, five of them are presented in the course of this thesis, which originally stem from Kratz et al. (2018), Acerbi and Szekely (2014), Bayer and Dimitriadis (2019), Costanzino and Curran (2015) and McNeil and Frey (2000). The main objective of this thesis was to outline and analyse one-sided, unconditional ES backtests, which take into account the practical aspects listed in Proposition 3.3. Therefore, some adjusted versions of the original backtesting approaches are proposed within chapter 4 of this thesis, which facilitate the application of the related methodologies in real-world scenarios. The conducted analysis is up to the best of my knowledge one of the most comprehensive evaluations of multiple ES backtests under one single framework.

In the empirical size and power simulations in chapter 5 of this thesis, the Z-test from Costanzino and Curran (2015) as well as the Test 2 from Acerbi and Szekely (2014) disclosed some potential miss-alignments in their underlying framework. Furthermore, the intercept ESR backtest motivated by Bayer and Dimitriadis (2019) might be the most challenging approach for any practical implementation due to a rather high degree of complexity and computational effort. Nevertheless, all ES backtests in scope proofed to be suitable in realistic backtesting scenarios, as outlined in chapter 6 of this thesis. With respect to the traffic light system required by the Basel Committee, all conducted ES backtests displayed reasonable results. Overall, especially

the de-facto one sided multinomial backtest based on Kratz et al. (2018) and the combined ES residual backtest motivated by McNeil and Frey (2000) need to be highlighted, due to their high practical relevance and the excellent results they exhibit throughout the course of this thesis.

Nevertheless, this thesis also revealed some open issues which are still subject to further research. As an example, substantial differences in the achieved p-values can be observed across the applied backtesting methodologies in certain scenarios within the preceding chapter. Whereas, this has no major impact on the assigned traffic lights, it might become more relevant if estimation models ought to be categorized on a finer grid. Furthermore, the choice of the respective backtest might also become relevant when rather extreme backtesting significance levels like $\kappa = 0.0001$ are evaluated, as bootstrap decisions might be subject to potential changes in the test outcome, due to the stochastic nature of the applied simulation decision. Thus, although all considered ES backtests exhibit overall reasonable results, it is still not clear which approach is the most appropriate one, as minor differences in the test outcomes might occur depending on the underlying setting. In line with other contributions, the power analysis within this thesis revealed that a backtesting horizon of $T = 250$ observations, which is currently stipulated by the Basel Committee, is not sufficient in order to consistently detect certain levels of risk underestimation in ES forecasting models. Within the preceding chapter, a backtesting horizon of $T = 500$ observations was applied, which appears to yield a good trade-off between backtesting power and data intensity. Nevertheless, some further considerations might be taken into account to determine an optimal backtesting time frame for the application of ES backtests. As one further aspect, there are still few contributions, like Du and Escanciano (2017), which focus on the development of a conditional ES backtest. Although this might not be of particular interest for practical applications, this is still a relevant aspect from a theoretical point of view.

Concluding this thesis, there is definitely further research that needs to be done in order to agree on an industry standard to backtest the ES. Nevertheless, I believe that difficulties in backtesting the ES should not any more be seen as a legit argument to favour the VaR as a primary risk measure for market risk, based on the results outlined within this thesis.

References

Acerbi, C. (2002). Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking & Finance*, *26*(7), 1505 - 1518.

Acerbi, C., & Szekely, B. (2014). Back-testing expected shortfall. *Risk Magazine*.

Acerbi, C., & Tasche, D. (2002). On the coherence of expected shortfall. *Journal of Banking & Finance*, *26*(7), 1487 - 1503.

Angelidis, T., Benos, A., & Degiannakis, S. (2004). The use of GARCH models in VaR estimation. *Statistical Methodology*, *1*(1), 105 - 128. doi: https://doi.org/10.1016/j.stamet.2004.08.00

Artzner, P., Delbaen, F., Eber, J.-M., & Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, *9*(3), 203-228. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9965.00068 doi: 10.1111/1467-9965.00068

Azzalini, A., & Valle, A. D. (1996, 12). The multivariate skew-normal distribution. *Biometrika*, *83*(4), 715-726. Retrieved from https://doi.org/10.1093/biomet/83.4.715 doi: 10.1093/biomet/83.4.715

Basel Committee. (n.d.-a). Fundamental review of the trading book: A revised market risk framework.

Basel Committee. (n.d.-b). Minimum capital requirements for market risk.

Basel Committee. (n.d.-c). Overview of the amendment to the capital accord to incorporate market risks.

Basel Committee. (n.d.-d). Supervisory framework for the use of backtesting in conjunction with the internal models approach to market risk capital requirements.

Bayer, S., & Dimitriadis, T. (2019). *Regression based expected shortfall backtesting.* (Working paper)

Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. *Journal of Business & Economic Statistics*, *19*(4), 465–474. Retrieved from http://www.jstor.org/stable/1392281

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *31*(3), 307 - 327. Retrieved from http://www.sciencedirect.com/science/article/pii/0304407686900631 doi: https://doi.org/10.1016/0304-4076(86)90063-1

Carver, L. (2013). Mooted VaR substitute cannot be backtested, says top quant. *Risk*, *26*.

Chen, J. M. (2014). Measuring market risks under the Basel Accord: VaR, stressed VaR and expected shortfall. *Aestimatio: IEB international Journal of Finance*, *8*, 184-201.

Christoffersen, P. (1998). Evaluating interval forecasts. *International Economic Review*, *39*(4), 841-62. Retrieved from https://EconPapers.repec.org/RePEc:ier:iecrev:v:39:y:1998:i:4:p:841-62

Cont, R., Deguest, R., & Scandolo, G. (2008, 01). Robustness and sensitivity analysis of risk measurement procedures. *Quantitative Finance*, *10*, 593-606. doi: 10.2139/ssrn.1086698

Costanzino, N., & Curran, M. (2015). Backtesting general spectral risk measures with application to expected shortfall. *Journal of Risk Model Validation*, *9*, 21-31. doi: 10.21314/JRMV.2015.131

Costanzino, N., & Curran, M. (2018, 01). A simple traffic light approach to backtesting expected shortfall. *Risks*, *6*, 2. doi: 10.3390/risks6010002

Danielsson, J., Embrechts, P., Goodhart, C., Keating, C., Muennich, F., Renault, O., & Shin, H. (2001, 08). An academic response to Basel II. *Financial Markets Group Special Papers*(130).

Diebold, F. X., Gunther, T. A., & Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, *39*(4), 863–883. Retrieved from http://www.jstor.org/stable/2527342

Dimitriadis, T., & Bayer, S. (2019). A joint quantile and expected shortfall regression framework. *Electronic Journal of Statistics*, *13*(1), 1823–1871.

Du, Z., & Escanciano, J. C. (2017). Backtesting expected shortfall: Accounting for tail risk. *Management Science*, *63*(4), 940-958. Retrieved from https://doi.org/10.1287/mnsc.2015.2342 doi: 10.1287/mnsc.2015.2342

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Chapman & Hall.

Emmer, S., Kratz, M., & Tasche, D. (2015). What is the best risk measure in practice? a comparison of standard measures. *Journal of Risk*, *18*, 31-60. doi: 10.21314/JOR.2015.318

Fissler, T., & Ziegel, J. F. (2016). Higher order elicitability and osband's principle. *Annals of Statistics*, *44*(4), 1680–1707.

Foellmer, H., & Schied, A. (2002, 09). Convex measures of risk and trading constraints. *Finance and Stochastics*, *6*, 429-447. doi: 10.1007/s007800200072

Gneiting, T. (n.d.). Making and evaluating point forecasts. , *106*, 746-762.

Hull, J. C. (2015). *Risk management and financial institutions* (Fourth Edition ed.). Wiley.

JP Morgan. (1996). *Riskmetrics, t.m.* (Tech. Rep.).

Kerkhof, J., & Melenberg, B. (2004). Backtesting for risk-based regulatory capital. *Journal of Banking & Finance*, *28*(8), 1845-1865.

Kratz, M., Lok, Y. H., & McNeil, A. J. (2018). Multinomial VaR backtests: A simple implicit approach to backtesting expected shortfall. *Journal of Banking & Finance*, *88*(C), 393-407. Retrieved from https://ideas.repec.org/a/eee/jbfina/v88y2018icp393-407.html doi: 10.1016/j.jbankfin.2018.0

Kupiec, P. H. (1995). Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives*, *3*, 73-84.

McNeil, A. J., & Frey, R. (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. *Journal of Empirical Finance*, *7*(3-4), 271-300. Retrieved from https://EconPapers.repec.org/RePEc:eee:empfin:v:7:y:2000:i:3-4:p:271-300

McNeil, A. J., Frey, R., & Embrechts, P. (2015). *Quantitative risk management: Concepts, techniques and tools* (Revised Edition ed.). Princeton University Press.

Moldenhauer, F., & Pitera, M. (2018, 08). Backtesting expected shortfall: a simple recipe? *SSRN Electronic Journal*.

Nass, C. A. G. (1959). The x^2 test for small expectations in contingency tables, with special reference to accidents and absenteeism. *Biometrika*, *46*(3-4), 365-385. Retrieved from https://app.dimensions.ai/details/publication/pub.1059416861 doi: 10.1093/biomet/46.3-4.365

Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, *59*(2), 347–370.

Nolde, N., & Ziegel, J. F. (2017). Elicitability and backtesting: Perspectives for banking regulation. *Annals of Applied Statistics*, *11*(4), 1833–1874. Retrieved from https://doi.org/10.1214/17-AOAS1041 doi: 10.1214/17-AOAS1041

Osband, K. (1985). *Providing incentives for better cost forecasting* (Unpublished doctoral dissertation). University of California.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5*, *50*(302), 157-175. doi: 10.1080/14786440009463897

Saerens, M. (2000). Building cost functions minimizing to some summary statistics. *IEEE transactions on neural networks*, *11*, 1263-71.

Thomson, W. (1979). Eliciting production possibilities from a well-informed manager. *Journal of Economic Theory*, *20*(3), 360 - 380.

Wong, W. K. (2008). Backtesting trading risk of commercial banks using expected shortfall. *Journal of Banking & Finance*, *32*(7), 1404 - 1415. Retrieved from http://www.sciencedirect.com/science/article/pii/S0378426607003627 doi: https://doi.org/10.1016/j.jbankfin.2007.11.012

Junior Management Science 6(3) (2021) 637-672

Junior Management Science

journal homepage: www.jums.academy

Talking is Silver, Doing is Gold? – The Influence of Corporate Social Responsibility on Corporate Financial Performance

Victoria Roth

Ludwig-Maximilians-Universität München

Abstract

Sustainability has become an omnipresent topic in the media and public as well as private debate. Stakeholders see the responsibility to promote sustainability with companies, pressuring them to increase their Corporate Social Responsibility (CSR). The relationship between CSR, being a means to satisfy a firm's stakeholders, and Corporate Financial Performance (CFP) is extensively debated in academics. This thesis contributes to this debate and tries to overcome measuring inaccuracies of previous studies by strictly categorizing CSR activities into CSR communication and CSR commitment. A total of 656 annual and CSR reports were examined, and variables representing these CSR activities were carefully and elaborately created, resulting in, among other things, a CSR communication breadth index, as well as an accurate assessment of communication quantity. A panel data analysis on European firms across industries over the observation period of eight years was conducted. The results reveal that only CSR communication has an influence on CFP. While standalone CSR reports and communication breadth have a positive influence, high levels of communication quantity have a negative effect.

Keywords: Corporate Social Responsibility; corporate financial performance; CSR commitment; CSR communication; stakeholder theory.

1. Introduction

1.1. Topic, relevance and research question

Microsoft aims to be carbon negative by the year 2030[1], BlackRock has a $60 billion platform of dedicated sustainable investment solutions[2], and McDonald's has done the previously unthinkable and launched a "McVegan"[3]. These are just a few examples of sustainable commitments that firms have communicated recently. Sustainability issues like climate change, the energy system transformation, gender equality, or the overexploitation of natural resources are gaining more and more public attention (Rockstrm et al., 2009; Rosen & Sellers, 1999). The responsibility of promoting sustainability lies no longer only with politics, but also the business world. The actions of companies have, therefore, increasingly attracted the attention of stakeholder groups (Joyner & Payne, 2002). Companies face high pressure from various stakeholders to conduct their business in an ethical manner (Elias, 2004). This pressure has made it clear to companies that they must adapt their practices to keep their stakeholders satisfied (Muller & Kräussl, 2011). Many companies recognize this necessity and have increased their commitment to CSR (Shahzad & Sharfman, 2017). Stakeholders are not only demanding more commitment to CSR but are also increasingly asking for companies to provide more information about how they are managing their CSR commitments (Sroufe & Gopalakrishna-Remani, 2019). The annual reporting of exclusively financial information does no longer suffice for stakeholders, and they expect firms to also disclose detailed non-financial information about the firm's activities and operations (Bernardi & Stark, 2018; Cormier & Magnan, 2014; Haque, 2017). Firms have met this growing demand for information on CSR activities by increasingly conducting and publishing CSR reports (Berliner & Prakash, 2015; García-Sánchez, Hussain, Martínez-Ferrero, & Ruiz-Barbadillo, 2019).

Both the increased commitment to CSR and the reporting

[1] https://news.microsoft.com/climate/

[2] https://www.blackrock.com/corporate/sustainability

[3] https://www.mcdonalds.com/de/de-de/product/big-vegan-ts-2560.html

DOI: https://doi.org/10.5282/jums/v6i3pp637-672

on it comes at a considerable price for companies, as these CSR activities are time and resource consuming. It would, therefore, seem as if CSR is merely a significant cost factor that companies accept to avoid the anger of stakeholders (Chiu & Sharfman, 2011). But could there be more to it than just bowing to social pressure? Might there be an economic interest in the engagement in CSR activities for firms?

For decades now, the relationship between CSR and CFP has received considerable academic attention. However, the results of the numerous studies addressing this topic are still inconclusive (Baboukardos, 2018; Cahan, De Villiers, Jeter, Naiker, & Van Staden, 2016; Cahan et al., 2016; Clarkson, Fang, Li, & Richardson, 2013). Meta-analyses of earlier research found that there is an overall positive relationship between CSR and CFP (Margolis & Walsh, 2003; Orlitzky, Schmidt, & Rynes, 2003), however, there is still an overwhelming amount of studies whose results differ considerably. Consequently, the debate on the link between CSR and CFP is far from over. (Shahzad & Sharfman, 2017) Research supporting the notion of a positive relationship between CSR and CFP argues that CSR is a way of managing and, therefore, improving a firm's relationships with its stakeholders. These relationships are said to positively influence CFP, mainly through the generation of a trustworthy image and a good reputation. A good reputation among many different stakeholder groups can positively influence CFP in various ways. Meeting the ethical standards of customers can improve brand reputation, which will subsequently increase brand loyalty and sales (Bird, Hall, Momentè, & Reggiani, 2007; Singal, 2014). Research has also shown that good corporate reputation attracts more talent and boosts employee morale, which, in turn, improves productivity (Coldwell, Billsberry, Van Meurs, & Marsh, 2008; Shafer, Smith, & Linder, 2005). Upholding a high moral standard also lowers the risks of lawsuits being filed against a firm, the implementation of additional regulations, and, therefore, the related costs (Karpoff, Lott, & Wehrly, 2005; Russo & Fouts, 1997). Research has also shown that long-term CSR provides insurance-like benefits to companies in and after times of crisis (Shiu & Yang, 2017). Neglecting the relationships with stakeholders, on the other hand, can have serious consequences for firms, as stakeholders provide resources to the company which influence its economic viability and efficiency (Donaldson & Preston, 1995).

CSR reporting is a way for companies to report on their CSR commitment to their stakeholders in a targeted and controlled manner. In doing so, the demands of the stakeholders can be addressed in a manner that presents the company in a favorable way. CSR reporting also helps a firm establish awareness and a reputation on the stock market (Liesen, Figge, Hoepner, & Patten, 2017; L. Wang & Tuttle, 2014). Analysts appreciate CSR disclosures because they provide more transparency and reduce the information asymmetry between companies and financial market participants (Bernardi & Stark, 2018; Dhaliwal, Radhakrishnan, Tsang, & Yang, 2012). Publishing CSR information, therefore, attracts the attention of analysts, thus, making the firm's stocks a more attractive investment to the other market participants. Transparent companies are also considered to be a less risk-prone, making them a preferred investment option for many (Alexander & Buchholz, 1978; Spicer, 1978).

However, the increase in CSR reporting does not always promote transparency. Some researchers have criticized that CSR commitment does not always keep pace with the increase in CSR communication, resulting in a communication-behavior gap. If stakeholders recognize this discrepancy, it can lead to a loss of reputation for the company. This gap does not necessarily have to be consciously created by the company. However, since CSR reporting is one way of improving the company's image to the outside world, this gap can also be a calculated strategy for concealing low-level CSR commitment. In both cases, however, it is clear that CSR commitment and CSR communication should not be considered to be the same thing.

The present work addresses some of the weaknesses of previous research and tries to overcome measurement inaccuracies of such intangible activities like CSR. Previous research has not made a clear distinction between CSR commitment and CSR communication so that the evaluation of CSR commitment has often been based on CSR communication. This thesis emphasizes the clear differentiation of both activities and treats them separately. Thus, CSR commitment is only considered as such if it has been evaluated independently from external resources.

This thesis aims to shed light on the influence of both CSR communication and CSR commitment on the financial performance of companies. Therefore, the following research questions will be addressed in this work:

Do CSR commitment and CSR communication separately influence CFP, and do these effects differ in their direction and strength? Or simply put, does it really pay to be good, or does a company just have to present itself well?

To investigate these research questions, an analysis with a complex data collection from 656 annual and CSR reports of 49 European companies from various industries over an observation period of eight years was performed. Based on this data, clearly defined variables were constructed. Hence, CSR communication was not only measured by the quantity of reporting but also by the breadth of the content of this communication. To ensure accuracy, the exact number of words of the CSR reporting of each company for each year was worked out.

In addition, the main part of this work focused on developing an index to measure the breadth of a company's CSR communication. Little existing research has dealt with the content of CSR reporting in such depth.

Furthermore, CFP was evaluated from an accounting as well as a stock market perspective. Because of this, shareholders were examined separately, and also distinguished from the rest of the stakeholders in this thesis, even though they technically belong to the groups of stakeholders.

The accurate and specific definition, creation, and use of these variables represents the impactful contribution of this thesis to present research in the field of CSR. The results offer

not only important theoretical implications but also reveal practical implications, valuable to firms that are attentive to their stakeholders' moral attitudes.

During the implementation of the main study, it became apparent that further relevant and interesting information could be extracted from the data. Therefore, two followup studies were conducted. The first followup study aimed to clarify whether the communication of different CSR topics also has a different influence on CFP. The second followup study specifically addressed the problem of the communication-behavior gap. The aim was to find out whether certain companies show discrepancies between their CSR commitment and their CSR communication and whether, and if so, how this has an impact on CFP.

1.2. Structure and chapter outline

At the beginning of the following chapter, essential concepts of CSR are defined. This is followed by the introduction of stakeholder theory as the theoretical basis of this thesis. Subsequently, the connection between this theory, CSR, and CFP is described. In this course, the hypotheses are derived. Afterwards, the risk of a communication-behavior gap, arising from discrepancies between the levels of CSR commitment and CSR communication, is addressed. The third chapter deals with the data used for the subsequent regression analysis. It is explained where and how the data was collected and how the final sample was obtained. This is followed by a description of how the dependent, independent, and control variables were created, and why these specific variables were chosen. Subsequently, the choice of the models for the regression analysis is explained. In chapter four, the regression results are presented. As the potential of the available data did not yet appear to be satisfactorily exhausted, two followup studies were conducted. The method behind these studies and their results are presented in chapter five. The results of the main study, as well as the followup studies, are then extensively discussed in the sixth chapter. In this general discussion, explanations for the findings of the studies are provided. This work concludes with its limitations, which simultaneously present opportunities for future research, and some theoretical as well as practical implications. The final chapter of this thesis contains a brief conclusion of all of the preceding, emphasizing some final and reflected thoughts.

2. Theoretical foundation and hypotheses

In order to get a comprehensive overview of the relationship between CSR and CFP, the following chapter provides a broad theoretical background. The chapter starts with the definition of the most important concepts. In the next step, the theoretical basis of this thesis, the stakeholder theory, will be explained. Subsequently, and on the basis of this theory, the relationship between CSR and CFP will be discussed. During this course, the hypotheses of this thesis will be derived.

2.1. CSR concepts and definitions

2.1.1. Sustainability

To set a basis for the following definitions, it is essential that the concept of "sustainability" is explained first. This concept was first introduced in "The World Conservation Strategy" report in 1980 (IUCN, UNEP, & WWF, 1980). In 1987, it was further debated in the Brundtland Report at the World Commission on the Environment of the United Nations (UN), where it was described as "development that meets the needs of the present generation, without compromising the ability of future generations to meet their own needs" (United Nations, 1987, p. 13). In the field of sustainable development, this conceptualization of sustainability has received wide recognition and application. Since the Brundtland Report was released, many new definitions of sustainability have emerged and must number in the hundreds (Vos, 2007). A term that has received great attention and which conceptually seems to be the basis of most definitions of sustainability is the „Triple Bottom Line" (Elkington, 2004). This framework represents the three pillars of sustainability: economic, social, and environmental considerations (Barkemeyer, Holt, Preuss, & Tsang, 2014). The three elements form a relationship should reinforce and support one another (Vos, 2007). Often, the social dimensions of sustainability have not been given the same importance as economic benefits (Remmen, 2007). The intangible nature of ethical and social benefits might be an explanation for this behavior (Sroufe & Gopalakrishna-Remani, 2019). It has been argued that social and environmental sustainability cannot be separated (Brundtland, 1987).

2.1.2. Sustainable Development Goals

In order to promote sustainability in developing and industrial countries, the General Assembly of the UN developed a new Agenda for Sustainable Development, the Sustainable Development Goals (SDGs), which consist of 17 goals and 169 targets and were agreed upon and passed in 2015. The SDGs, much like their predecessor the Millennium Development, are a universal set of goals, indicators, and targets which were introduced at the Rio +20 Conference. The SDGs are building on the Millennium Development Goals (MDGs), which were set in 2000. The SDGs expanded the MDGs' agenda and included further issues such as sustainable consumption, innovation, and climate change. This set of goals is at the heart of the 2030 Agenda, acting as a universal agenda and serving as "a shared blueprint for peace and prosperity for people and the planet, now and into the future" for UN member states to use to frame their agendas and policies (United Nations, n.d.). These goals aim to end poverty, improve health and education, reduce inequality, and increase economic growth while preserving the world's oceans and forests to tackle climate change (United Nations, 2016). The overall goal of this framework is to ensure prosperity for all and to protect the planet for future generations. The governments, businesses, and civil society of all countries are required to actively work on achieving these goals by the year

2030 (Hoque, Rahman, Molla, Noman, & Bhuiyan, 2018). Businesses are one of the stakeholder groups this agenda is aimed at. Their proactive engagement is needed to attain the SDGs, and they have been urged to incorporate these targets into their CSR strategy (García-Sánchez et al., 2019).

2.1.3. CSR

There is an abundance of definitions of CSR whose content has continuously evolved (Argandoña & von Weltzien Hoivik, 2009; Maak, 2008). The fact that the dynamics of the discussion on CSR are permanently influenced by the continually changing topics that are of particular relevance to society makes it all the more difficult to grasp a static definition of the term. Many of the existing definitions agree on certain points regarding the core features of CSR. CSR involves voluntary actions to promote a social good, which are not just following existing regulations in the countries the firm operates in but go beyond the interests of a company and legal requirements (McWilliams & Siegel, 2000; Rodriguez-Fernandez, 2016). It is about companies incorporating economic, social, and environmental impacts into their operations. This can include a variety of adjustments like environmentally friendly production, employee satisfaction, respect for local communities, consideration towards investors, and general ethical behavior (Bénabou & Tirole, 2010). The Commission of the European Communities, therefore, very fittingly defines CSR as "a concept whereby companies integrate social and environmental concerns in their business operations and in their interaction with their stakeholders on a voluntary basis" (Commission of the European Communities, 2002, p. 3).

2.1.4. CSR reporting

A firm's reports containing its sustainability-related information can take many names, such as "CSR reports", "Sustainability Reports", or "Citizenship Reports" (Roca & Searcy, 2012). Not only the name but also the definitions for the concept of CSR reporting vary considerably. The definition of the World Business Council for Sustainable Development (WBCSD) has often been used as a representative example. They define CSR reports as "public reports by companies to provide internal and external stakeholders with a picture of the corporate position and activities on economic, environmental and social dimensions" (WBCSD, 2003). CSR reports are, therefore, non-financial reports containing information on the economic, social, and environmental impacts of a firm's operations. This sets them apart from environmental reports, which, as the name suggests, only contain information on the actions of companies that are in some form related to the environment. CSR reports provide a more comprehensive overview of the non-financial aspects of a company's managerial strategies and decisions, giving information on the company's values, governance, and social goals. (Berthelot, Coulmont, & Serret, 2012) These reports can either be integrated publications or take the form of a standalone report (Daub, 2007; Schaltegger, Bennett, & Burritt, 2006).

As CSR reporting is voluntary in most jurisdictions, mandatory requirements regarding these publications are mostly absent. The increased use of CSR reports made researchers aware that the development of standards and guidelines for the creation of such reports is essential. Corporations have, therefore started to use reporting guidelines, like the guidelines created by the Global Reporting Initiative (GRI) (GRI, 2011), as an orientation in the development of their CSR reports. These guidelines enhance the comparability of CSR reports and are a way of reducing the cost of conducting the reports (Berthelot et al., 2012). Considering that the use of these guidelines is voluntary as well, it seems like firms have a lot of freedom when it comes to creating their CSR reports (Searcy & Buslovich, 2014). Some progress has been made concerning the regulation of non-financial disclosure. As a result of the EU's initiative on CSR, the 2013 Accounting Directive was amended. Companies with more than 500 employees or a public listing are required to disclose non-financial information on environmental, social, and employee matters as well as their respect for human rights, anti-corruption, bribery, and corporate diversity. This only started to be mandatory for reports published in 2018, covering the fiscal year of 2017-2018. (Cheng, Ioannou, & Serafeim, 2014; Schulz & Kourkoulas, 2014)

2.2. CSR and CFP

2.2.1. Shareholder theory

After reviewing the extant literature, it became apparent that stakeholder theory is the most appropriate theory to provide a rationale to the research question of this thesis. The concept of stakeholder theory has been brought forward by Freeman (Freeman, 1984). Stakeholder theory, as an integrative and holistic perspective, has become the most prominent theoretical foundation in CSR research (Carroll, Brown, & Buchholtz, 2017; Montiel & Delgado-Ceballos, 2014; Perrini & Tencati, 2006; Weber & Marley, 2012). Conducting business does not simply take place between two single actors. A lot of additional stakeholders are involved who play crucial roles in a firm's success (Schaltegger, Hörisch, & Freeman, 2019). Freeman defines stakeholders as "any group or individual who can affect or is affected by the achievement of the organization's objectives" (Freeman, 1984, p. 46). Ackoff (1974) adds the notion of legitimacy and clarifies that stakeholders are a group that the firm needs in order to exist. The underlying premises of stakeholder theory are rather intuitive. A company is in relationships with many stakeholder groups, and a firm has obligations to the members of these groups. The stakeholder theory differs from traditional theories in the sense that it takes the view that financial value creation is not the sole focus of a firm. Therefore, while stakeholder theory still considers shareholders to be one of those interest groups, it does not consider them to be the most important one (Hörisch, Freeman, & Schaltegger, 2014). The stakeholder theory views a firm's relationships with all its stakeholders as a crucial component of its management strategy and believes that those relationships should

Figure 1: Overview of the Sustainable Development Goals, Source: United Nations

be considered in important managerial decisions (Searcy & Buslovich, 2014). These relationships with stakeholders are considered to be crucial to the functioning of a firm (Freeman, 1984, 2010), and in order to understand how a company functions, it is important to understand how these relationships work and change. Achieving and pertaining to this understanding is the job of the executives. They have to manage and shape these relationships in a way that creates the most value for their company (Freeman, 1984). It would be impossible to operate a business if the relationships with internal and external stakeholders were not robust (Freudenreich, Lüdeke-Freund, & Schaltegger, 2019). According to stakeholder theory, all stakeholders have to be acknowledged and satisfied as the dissatisfaction of any stakeholder group can hurt a firm's legitimacy in society and, consequently, its survival (Clarkson, 1995; Freeman, 1984).

According to Donaldson and Preston (1995), all stakeholder theory research can be classified based on three theoretical branches: descriptive, normative, and instrumental. The descriptive branch of stakeholder theory illustrates the nature of firms, as well as how their managers behave (Donaldson & Preston, 1995). The normative branch deals with the morality of the behavior of firms and their managers. The instrumental branch investigates the influence of the behavior of firms and managers on a firm's performance (T. M. Jones, 1995). To summarize, the three branches address the following questions: What are firms doing with respect to stakeholder relationships? What should firms do when dealing with their stakeholders? How does a firm's adherence to stakeholder management principles affect its performance? (T. M. Jones, 1995; Laplume, Sonpar, & Litz, 2008)

Although these branches of stakeholder theory are of a complementary nature (Donaldson, 1999), instrumental stakeholder theory provides the best theoretical basis for this thesis. The instrumental branch of stakeholder theory combines the stakeholder concept with economic theory, ethics, and findings from behavioral science (T. M. Jones, 1995). It highlights the relationship between stakeholder management and a corporation's traditional economic objectives, like CFP. Manager must fulfill the interests of stakeholders, economic or otherwise, in order to encourage them to participate in activities which create value for the corporation. (Laplume et al., 2008) Therefore, a connection between stakeholder management and CFP is assumed in research. In their influential article on this theory, Donaldson and Preston (1995) argue that corporations whose managers successfully establish stakeholder practices and principles and therefore manage their relationships with their stakeholders properly can achieve superior CFP over time (Donaldson & Preston, 1995; Freeman, 1984). Successfully managing stakeholder relationships may, therefore, result in competitive advantages (Barnett & Salomon, 2006). These relationships a firm has with its stakeholders have been described metaphorically as "contracts" (T. M. Jones, 1995). A firm could, therefore, be considered as a „nexus of contracts between the firm and its stakeholders (Jensen & Meckling, 1976). These contracts, like the contract between a firm and the community it is placed in, are not specific, formal, or in any way described in a document. Much like with real people, a firm is seen as a more desirable contracting partner when it shows a strong moral fiber, or at least shares the same moral values as its potential contracting partner. A way for stakeholders to get an idea of the moral tendencies of firms is the firm's reputation. Further, its policies and decisions, as well as the way it is directly dealing with its stakeholders, reflect this (Frank,

1988). Some of these policies and decisions are more visible to stakeholders than others, depending on if they directly or indirectly affect the stakeholders. Needless to say, all of them will influence the firm's reputation among its shareholders. Fostering a moral culture and acting morally as a firm can improve corporate reputation, as well as help restrain opportunistic behavior among employees. Therefore, it can reduce agency, as well as transaction costs. Therefore, good ethics do not hinder profitability, but can lead to more and better relationships and consequently to better business. (T. M. Jones, 1995)

2.2.2. CSR commitment and CFP

Instrumental stakeholder theory serves as an explanation of why companies feel the need to engage in CSR (Perrini & Tencati, 2006; Post, Preston, & Sachs, 2002; Van Marrewijk, 2003). Firms feel the growing pressure from the stakeholder side to act responsibly (Chiu & Sharfman, 2011). In order to maintain good relations with their stakeholders, companies must comply with this request. CSR, being time and resource consuming, comes at a financial cost for firms. According to instrumental stakeholder theory, profitability and CSR are not only not mutually exclusive, though. In order to achieve sustainable and superior CFP, firms have to align their corporate interests with appropriate CSR activities to satisfy their stakeholders (Davenport, 2000; Jain, Vyas, & Chalasani, 2016; Waddock & Graves, 1997). This point of view has sparked a long-lasting debate on the link between CSR and CFP, which has been the subject of academic research for decades now (Al-Tuwaijri, Christensen, & Hughes Ii, 2004; Anderson & Frankle, 1980; Belkaoui, 1976; Orlitzky et al., 2003). The results of these studies are contradictory, and there is much controversy surrounding the question of whether the correlation is positive, negative, or even existent. Some researchers find a negative relationship. They argue that CSR mainly produces additional costs, for example, from donating money to social causes like education, healthcare, and infrastructure, or from investing in environmentally friendly materials and machinery and customer service (Kapoor & Sandhu, 2010). Other studies (e.g., Abbott & Monsen, 1979; Griffin & Mahon, 1997; McWilliams & Siegel, 2000) have either found no or an inconclusive effect of CSR on CFP. The majority of studies, though, supports the hypothesis of a positive relationship between a company's CSR efforts and its financial performance (Flammer, 2015; Saeidi, Sofian, Saeidi, Saeidi, & Saaeidi, 2015). Therefore, this thesis will look at the relationship between CSR and CFP through the lens of instrumental stakeholder theory, arguing that CSR is an instrument used by companies to secure resources and support from stakeholder groups (T. M. Jones, 1995).

There are various ways in which this improved relationship with stakeholders can become apparent. Many of them can be associated with an improved reputation of the firm. Corporate reputation reflects the perception that stakeholders have of a company (Brammer & Pavelin, 2006). Showing a high degree of CSR commitment is a way of improving this reputation. Employees' knowledge capital, capabilities, and motivation are critical aspects that are engaged in the value creation in a company (Shafer et al., 2005). Studies have shown that a good corporate reputation can boost employee morale, leading to improved productivity and, consequently, better CFP (Basu & Palazzo, 2008; Brammer & Pavelin, 2006; Orlitzky et al., 2003; Pirsch, Gupta, & Grau, 2007). Companies with a higher sustainability reputation also attract and retain more talented employees as they feel like the companies match their ethical expectations better (Coldwell et al., 2008). Employing high skill workers is also likely to reduce the cost of quality management and the cost of production, which in turn will lead to improved CFP (Brown, Willis, & Prussia, 2000; Brown, 1996). Additionally, it has been shown that employees may even forgo financial benefits to work for a company with a superior CSR reputation (Sroufe & Gopalakrishna-Remani, 2019). Improving a company's social aspects like employee satisfaction, retention, and recruitment can minimize intra-firm conflicts between the duties and targets of different branches, departments, managers, shareholders as well as internal stakeholders (Wagner, 2007). Environmental and social philanthropy generates goodwill among stakeholders through positive publicity, resulting in a favorable public image (Ameer & Othman, 2012). This enhances loyalty among customers, as customers prefer the products of companies they deem conscious about their social and environmental responsibility (Creyer, 1997; Jacobs, Singhal, & Subramanian, 2010; Mohr & Webb, 2005). The projection of a socially responsible image also gives consumers the feeling of being able to evaluate a company better. At the same time, a good image also favorably affects the evaluation of products and services received (Rodriguez-Fernandez, 2016). Customers will buy a product if they feel like it meets their personal needs. This value of a product can also be symbolic (Freudenreich et al., 2019). CSR efforts can lead to better product image and, consequently, to higher sales and new market opportunities (Sroufe & Gopalakrishna-Remani, 2019). This superior corporate reputation can also prove to be valuable to a firm in case of crises, as it can help to protect the company against damage from possible negative future events (Bansal & Clelland, 2004; Godfrey, 2005; Godfrey, Merrill, & Hansen, 2009; Sharfman & Fernando, 2008).

Effective environmental management can have cost avoidance benefits as it can mitigate the risks of crises or regulation and the expenses associated with them (Reinhardt, 1999), as well as prevent the costs of lawsuits and legal settlements (Karpoff et al., 2005). The disregard of social responsibilities, however, might bring about further legislation, which can lead to higher costs of compliance (Russo & Fouts, 1997).

All of these described consequences of improved reputation among stakeholders can lead to higher CFP. In order for stakeholders to recognize real CSR commitment as such, it has to be externally and independently verified. Companies that can present such verification should benefit from these reputational advantages. Based on this, the following

hypotheses were derived:

> H_{1a}: Being awarded a high CSR score has a positive influence on accounting-based CFP.
>
> H_{1b}: Being awarded a sustainability award has a positive influence on accounting-based CFP.

Shareholders also seem to appreciate high levels of CSR commitment, as they can also be linked to the better performance of a firm on the stock market. Borgers, Derwall, Koedijk, and Ter Horst (2015) showed that the stocks of companies with environmentally friendly practices generated higher returns than stocks of companies with a substandard environmental record. The social aspect of CSR seems to influence stock returns as well. Stocks of firms that scored high on diversity, community, human rights, and employee relations outperformed stocks of firms with low scores in this regard (Kempf & Osthoff, 2007).

In recent years, CSR has become an important indicator of business performance (T. Wang & Bansal, 2012). CSR is often seen as an indication of management skills by shareholders (Bowman & Haire, 1975). Displaying concern and awareness for social and environmental issues might be regarded as a skill that enables managers to run a functioning and maybe even superior company, making the company an attractive investment (Alexander & Buchholz, 1978). Investors are willing to pay a premium for the stock of companies that appear to be well-governed (Coombs & Holladay, 2011). External validation of this responsible management seems to be rewarded as Klassen and McLaughlin (1996) found that winning environmental awards entails positive stock returns for companies, whereas negative stock returns follow the occurrence of environmental crisis.

There has been a rise in socially responsible investments, which are a way of directing an investment fund in a way that aligns the financial objectives and the social commitments to social and environmental issues of investors. This trend began to take off in the 1970s, starting in the US, while slowly but steadily moving to Europe as well. (Haigh & Hazelton, 2004) Prominent fund managers have started to incorporate environmental, social, and governance (ESG) criteria into their portfolio selection methods (Oh, Park, & Ghauri, 2013). Financial decision-makers display strong preferences and a considerably higher willingness to pay for socially responsible investments. Also, this willingness to pay is much higher for investment products that have been certified to be responsible, compared to those with no certification. Investors who have indicated to follow certain psychological motives, norms, and values. Getting the feeling of doing good and serving a cause from making socially responsible prevails investors to forgo higher returns, which might have been otherwise achieved through the investment in unsustainable counterparts. (Gutsche & Ziegler, 2019)

In times of crisis, investors seem to feel safer with socially responsible investments. While the overall size of professionally managed assets stayed approximately the same during the last financial crisis from 2007 to 2010, SRI assets displayed a healthy growth (Oh et al., 2013). Being socially responsible has also shown to provide insurance-like benefits to a company when looking at the effects of legal actions that have been filed against it (Godfrey et al., 2009). A lack of CSR practices can discourage investors, as investing in such firms seems to be of higher risk (Alexander & Buchholz, 1978; Spicer, 1978).

Drawing from the presented findings in previous research, it can be argued that the external validation of high CSR commitment has a positive influence on the market-based financial performance of a firm. Consequently, two additional hypotheses are put forward:

> H_{2a}: Being awarded a high CSR score has a positive influence on market-based CFP.
>
> H_{2b}: Being awarded a sustainability award has a positive influence on market-based CFP.

2.2.3. CSR communication and CFP

There is not only an increase in the requests from stakeholders for more CSR commitment but also for more information about how companies manage their CSR (Sroufe & Gopalakrishna-Remani, 2019). Companies face pressure from various stakeholders to be transparent about their CSR commitment (Elias, 2004). It is expected of firms to disclose not only financial but also non-financial information about the firm's activities and operations (Cormier & Magnan, 2014; Haque, 2017). Consequently, many previously strictly financial reports now contain information about the CSR activities of the firm, or even standalone CSR reports are published (Rupley, Brown, & Marshall, 2012). Most of the largest companies in the world are reporting on their CSR commitments (KPMG, 2011). CSR reports are one of the most effective moralized communication tools (Castelló, Morsing, & Schultz, 2013; Hooghiemstra, 2000). They act as facilitators and help firms to respond to the demand for information from stakeholders. By fulfilling these expectations, firms can manage their perceived legitimacy (Hahn & Kühnen, 2013). Roberts (1992) described a significant relationship between the level of a company's disclosure levels and its economic performance. Similarly, Murray, Sinclair, Power, and Gray (2006) found a long-term relationship between high disclosure levels and economic performance within UK companies. CSR reporting of a firm is mostly aimed at gaining acceptance from its key stakeholders (Schaltegger et al., 2006). These disclosures can lead to customers choosing between what they perceive to be more or less socially and environmentally friendly products, or employees choosing organizations that seem to represent their values as their employer. This is why CSR disclosures can lead to a competitive advantage for companies that know how to construct them in their favor. (Matisoff, 2013) The debate about CSR has long been extended beyond environmental practices to a much wider variety of issues (Sroufe & Gopalakrishna-Remani, 2019). Therefore, in order to display a high degree of responsibility, a company must communicate on a wide range of CSR topics (Rodriguez-Fernandez, 2016).

Considering the findings of existing literature on stakeholders' perception of CSR reporting, visible, extensive, and broad CSR communication should have a positive influence on the financial performance of a company. Therefore, the following hypotheses were developed:

H_{3a}: Issuing a standalone CSR report has a positive influence on accounting-based CFP.

H_{3b}: A higher quantity of CSR reporting has a positive influence on accounting-based CFP.

H_{3c}: Addressing a broader range of CSR topics has a positive influence on accounting-based CFP.

This increasing trend of requests for CSR transparency can not only be observed from the stakeholders' side (Michelon, Rodrigue, & Trevisan, 2020). Financial market participants are on a constant quest for information that can be incorporated in their risk assessment of risks and decision-making processes (Malsch, 2013). Some studies find that analysts use CSR reports for preparing their recommendations, as this information reduces forecast dispersion (Cormier & Magnan, 2014; Dhaliwal et al., 2012; Ioannou & Serafeim, 2015). CSR disclosures enhance transparency and therefore reduce the information asymmetry between firms and their shareholders (Kolk & Perego, 2010). The availability of CSR information has also proven to improve analysts' forecast accuracy and to decrease forecast error (Bernardi & Stark, 2018).

CSR reporting also escalates awareness of the firm's existence in the financial market and among investors (Merton, 1986). CSR reporting helps in establishing a firm's reputation on the financial market (Liesen et al., 2017; L. Wang & Tuttle, 2014), as financial market participants use it to form their perception of the company in question (Cheng et al., 2014; Luo, Wang, Raithel, & Zheng, 2015; Sharfman & Fernando, 2008). Publishing CSR information attracts the attention of analysts, which will, in return, make the company more attractive to investors (Bernardi & Stark, 2018; Dhaliwal et al., 2012). Were all other factors to remain constant, it has been shown that investors would decide in favor of companies that supply information on their social responsibility (Rodriguez-Fernandez, 2016). There is a significant relationship between the degree of a company's CSR disclosure and the economic performance of a firm (Roberts, 1992). Studies (e.g., De Klerk, De Villiers, & Van Staden, 2015) have shown that higher levels of CSR disclosure are associated with higher share prices. Flammer (2015) found that investors are willing to pay more for the stocks of companies that displayed a higher degree of social responsibility. Investors have shown to accept lower expected returns on investments if those also pursue social objectives (Richardson & Welker, 2001).

In summary, the findings from the existing literature indicate that clearly visible, extensive, and broad CSR communication should also have a positive influence on the performance of a company on the stock market. Therefore, the following hypotheses are put forward:

H_{4a}: Issuing a standalone CSR report has a positive influence on market-based CFP.

H_{4b}: A higher quantity of CSR reporting has a positive influence on market-based CFP.

H_{4c}: Addressing a broader range of CSR topics has a positive influence on market-based CFP.

2.2.4. Risks of a communication-behavior gap

While companies face increased pressure from various stakeholder groups to act ethically and be transparent (Elias, 2004), the everincreasing amount of CSR information communicated has been a prominent target of criticism. Just as for any form of communication, credibility is essential for CSR reports (Jackob, 2008), and the lack of such credibility has been the root of this criticism (Dando & Swift, 2003). This is why CSR reports may not decrease perceived information asymmetry and facilitate dialogue with stakeholders but actually widen the credibility gap and therefore threaten a firm's legitimacy (Seele & Gatti, 2017). The increasing level of CSR disclosure is not always met with increased confidence but sometimes skepticism about the sincerity of the reported information (Dando & Swift, 2003; Doane & Potts, 2000). Companies are often accused of using CSR reporting to pursue calculated and strategic goals rather than to feel a real commitment to their stakeholders (Brown & Deegan, 1998; Deegan, Rankin, & Tobin, 2002). CSR can be used as impression management to boost a company's image (Hooghiemstra, 2000). External strategies can potentially give a corporation the appearance of being ethical, even if such a commitment does not exist (Laufer, 2003). Sometimes unsubstantiated disclosures are published in order to respond to pressures from the public or manage the public perception of the company (Adams, 2002; Hooks, Coy, & Davey, 2002). The content of CSR reports has also been harshly criticized as it seems like companies have too much freedom in choosing and altering the information they release (Coombs & Holladay, 2011; Fonseca, 2010). Firms can, therefore, choose which stakeholders to address and are able to deliver a positive and confident message to their target group (Fonseca, 2010; Milne & Gray, 2013). Inconsistencies between CSR reporting and a company's activities have made stakeholders al-most cynical towards CSR reporting and made them question the moral legitimacy of firms (Basu & Palazzo, 2008; Claasen & Roloff, 2012; Illia, Zyglidopoulos, Romenti, Rodríguez-Cánovas, & del Valle Brena, 2013). Academics have been concerned with the fact that reported performance and actual performance are not always the same (Morhardt, Baird, & Freeman, 2002). Table 1 offers an overview of research, which has concerned itself with the gap between a firm's CSR commitment and its CSR communication.

Many stakeholders are pleased with the increased supply of CSR reports, but some are also skeptical as this development gives opportunities for deception, such as greenwashing. This is owed to the issues with comparability of the reports and the fact that the regulatory progress of this report-

Table 1: Studies addressing the communication-behavior gap

Hammond and Miles (2004)	High quality CSR is not necessarily synonymous with responsible performance, and indeed research indicates this (p. 62).
Dando and Swift (2003)	Such a general lack of third-party assurance has likely contributed to general critique that social reporting is simply about corporate spin and public relations (p. 197).
Laufer (2003)	. . . the growing body of social and environmental accounting research that finds corporate posturing and deception in the absence of external monitoring and verification, e.g., the structuring of corporate disclosures so as to maximize perceptions of legitimacy (p. 253).
Lock and Seele (2017)	Credibility in organizational communication is often challenged because of discrepancies between words and deeds that result in "credibility" or "legitimacy" gaps (p. 585).
Siano, Vollero, Conte, and Amabile (2017)	Deceptive communication, often labelled as greenwashing, has become a recurrent practice in the context of marketing and corporate communica-tion strategies, aimed at hiding the most controversial aspects related to corporate sustainability. There are several possible reasons for this emerging trend, such as the increasing pressure on companies from dif-ferent stakeholder groups, and the potential benefits that can be gained by a company in terms of financial performance and reputational capital (p. 27).
Milne and Gray (2013)	What is also suggested here and has been of concern to academics for some time, is that reporting and performance are not the same (p. 9).

ing has been slower in its development than the increasing demand for this kind of information. (Delmas & Burbano, 2011)

Often, there is a mismatch between shareholders' expectations in firm behavior and their perception of what firms actually do. Discrepancies between words and actions in a firm's behavior can result in credibility gaps in their communication (Sethi, 1975). These gaps between firms and stakeholders can also occur when the firms do not communicate their CSR efforts in a credible way (Illia et al., 2013). This can compromise their legitimacy in society, jeopardizing their "license to operate" in society (Donaldson & Dunfee, 1999). It generally seems like there is a divergence between stakeholder's expectations of a company's CSR and the company's actions (Dando & Swift, 2003).

Firms from industries with negative associations such as oil or chemicals might engage in strategies that are supposed to shift the attention away from the firm's business and create confusion to enhance firm objectives and accomplishments in a deceptive way (Laufer, 2003). Among the top reporters are firms with the worst history of environmental impact, having spent a lot of resources to acquire a green image (Hammond & Miles, 2004). Publicizing high CSR performance might not be about gaining credibility but more about shifting the attention away from actually poor social and environmental performance (Hammond & Miles, 2004). Firms with poor social and environmental performance draw a bigger benefit from manipulating the image stakeholders have of the firm's CSR performance (Clarkson, Li, Richardson, & Vasvari, 2008). In this sense, companies with lower CSR performance should be expected to issue standalone CSR reports to deflect from their mishaps, improve their public reputation to appear as a good corporate citizen, and legitimize their standing with stakeholders (Mahoney, Thorne, Cecil, & LaGore, 2013). CSR disclosure could rather be seen as a means for mediating the effect of poor performance on a company's reputation (Cho, Roberts, & Patten, 2010).

Hall defined reputation as "a fragile resource; it takes time to create, it cannot be bought, and it can be damaged easily" (Hall, 1993). Companies are in constant interaction with their stakeholders. These interactions do not always have a positive outcome and put a company's reputation at risk (Fombrun, Gardberg, & Barnett, 2000). As previous research has suggested that there is a positive relationship between CSR initiatives and corporate reputation, managing these activities becomes extremely important (Bebbington, Larrinaga, & Moneva, 2008).

Greenwashing implies that the disclosed CSR information does not necessarily correspond with a firm's actual social and environmental performance (Cho & Patten, 2007). It refers to the fact that companies creatively manage and repair their public reputation with regard to environmental performance. They do so by issuing disinformation like deflecting attributions of fault, disguising the true nature of problems or accusations, hiding deviance, or reattributing blame. This is done to appear in a good light and some sort of leadership position (Laufer, 2003). Greenwashing describes the situation when companies are willing to take up the costs of voluntary disclosure to publish misleading or false CSR information to trick their stakeholders into believing that they are responsible (Cho et al., 2010; Guidry & Patten, 2010; Lyon & Maxwell, 2011). Greenwashing does not necessarily have to involve the disclosure of false information, but the conscious and advantageous omission of negative and dam-

aging information (Mahoney et al., 2013). Firms that engage in labor rights abuses would not disclose in such legal non-compliance to avoid reputational loss and possible restrictions (Belal & Cooper, 2011).

This communication-behavior gap clearly shows that CSR commitment and CSR communication should not be considered to be the same thing. This is why a clear distinction was made between CSR commitment and CSR communication in this thesis, and the influences of both types of CSR activities on CFP were examined.

2.3. Overview of hypothesis

Table 2 provides an overview of all hypotheses derived in the previous sections. A simplified illustration of the hypotheses can be found in the Appendix A1.

3. Data and Method

The biggest challenge of research in the field of CSR is the identification of tangible parameters that can be used to measure CSR activities. Therefore, much attention has been paid to the creation of the variables assessing CSR activities. The main contribution of this work was, therefore, the construction of some of the following variables. In order to transparently reflect the complex process of creating these variables, the method behind their construction is already explained in the definition of the variables themselves. First, the process of collecting the data, and its preparation for the statistical analysis is described. Afterward, the data sources are named, and the sample selection is explained. This is followed by the presentation of the dependent, independent, and control variables. The chapter concludes with the statistical method, explaining the choice of the regression models.

3.1. Data sources

The data on financial performance, as well as the number of employees, ESG scores, sustainability awards, industry membership, and firm nationality, were all collected from the Thomson Reuters Datastream. This platform contains financial and macroeconomic data for companies in 175 countries and 60 markets (Thomson Reuters, 2013).

In order to collect the data on CSR communication, the sampled firms' websites and the annual reports and CSR reports for the years 2010 to 2017 were downloaded in the English language. To ensure of the continuity and completeness of the data, the firms' archival data was complemented with CSR reports gathered from the GRI database. The GRI database contains CSR reports published from 1999 until the date for more than 6,000 companies worldwide (García-Sánchez et al., 2019). If there was no report on either website, it was assumed that no CSR report was published in the respective year. The information on a company's age was retrieved from their websites as well.

3.2. Sample selection

For increased comparability, only firms from European countries were selected. All of these countries are members of the UN. Therefore, the fact that all the firm's headquarters are located in Europe makes them more comparable with regard to regulations concerning social, environmental, and economic laws and regulations as well as CSR reporting standards and requirements. Strongly differing legal and regulatory frameworks may have affected both the level of CSR disclosure and CFP (Beck, Frost, & Jones, 2018). The observation period from 2011 to 2018 was deliberately chosen. First of all, 2018 was the most recent year for which uniform and complete business figures were available for most companies. In addition, a major environmental scandal, the Volkswagen emissions scandal, was uncovered in 2015. In this scandal, fraud in sustainability reporting was uncovered and punished on an unprecedented scale and may have sensitized companies to false or exaggerated reporting. The selected time frame can be divided into two roughly equally long periods before and after the exposure of the scandal. This scandal is by no means the subject of this paper. It only served as an orientation for the selection of the investigation period.

To create a first pool of potential firms, the European companies, which are in the 80th percentile of their industry in terms of market capitalization, were preselected from the Thomson Reuters database. This selection criterium was chosen to enhance the comparability of the companies among each other. As all of these firms are publicly listed, all of them and their activities are of public interest, giving them a certain degree of publicity. To make sure that sufficient historical data was available for the regression analysis, companies with less than eight years of continuous data on ESG scores, CFP, employee numbers, total liabilities, total assets, and sustainability awards were eliminated. In the next step, the annual and CSR reports for these companies were retrieved. If it was not possible to access all the annual reports of a company for the relevant years, or if it was evident that a CSR report was issued, but not accessible, the company was eliminated from the sample. This reduced the sample to 49 companies. A total of 656 reports (392 annual and 264 CSR reports) were collected.

3.3. Variables

3.3.1. Dependent variables

This study examines the influence of CSR activities on CFP. A variety of proxies for CFP have been used in previous research. While many studies have looked at accounting-based proxies like ROA and ROE, others have examined how stocks perform in relation to their CSR activities (Borgers et al., 2015; Luo et al., 2015). Often, studies only looked at either accounting-based or market-based CFP. This work, stakeholder reactions to CSR activities are being debated, but a special look is taken at shareholders, too. Therefore, both a proxy for market-based CFP and a proxy for accounting-based CFP were used as dependent variables.

Table 2: Hypotheses overview

CSR commitment
H_{1a}: Being awarded a high CSR score has a positive influence on accounting-based CFP.
H_{1b}: Being awarded a sustainability award has a positive influence on accounting-based CFP.
H_{2a}: Being awarded a high CSR score has a positive influence on market-based CFP.
H_{2b}: Being awarded a sustainability award has a positive influence on market-based CFP.
CSR disclosure
H_{3a}: Issuing a standalone CSR report has a positive influence on accounting-based CFP.
H_{3b}: A higher quantity of CSR reporting has a positive influence on accounting-based CFP.
H_{3c}: Addressing a broader range of CSR topics has a positive influence on accounting-based CFP.
H_{4a}: Issuing a standalone CSR report has a positive influence on market-based CFP.
H_{4b}: A higher quantity of CSR reporting has a positive influence on market-based CFP.
H_{4c}: Addressing a broader range of CSR topics has a positive influence on market-based CFP.

ROA: Return on assets (ROA) served as the proxy for accounting-based CFP in this study. Many studies that do not include shareholder perception have used ROA as a measure of CFP (Van der Laan, Van Ees, & Van Witteloostuijn, 2008), as ROA does not reflect external market responses but internal decision-making and managerial performance (Orlitzky et al., 2003). ROA is one of the most widely used and broadest measures of a firm's operating performance used in CSR studies to test the relationship between a firm's financial performance and its CSR activities (Borgers et al., 2015). In the past, this measure has been found to relate better with CSR than alternative measures (McGuire, Sundgren, & Schneeweis, 1988; Waddock & Graves, 1997). Therefore, it was considered an appropriate proxy for accounting-based CFP for this study.

ROA is equal to operating profit before depreciation and provisions divided by total assets (Rodriguez-Fernandez, 2016). It is the ratio of net income to the book value of the total assets of a firm (Brigham & Ehrhardt, 2013) and provides information about the degree to which the revenues of a firm exceed its costs (Firer & Williams, 2003). If, for example, sales increase while the costs structure and assets remain constant, ROA will grow (Zellweger & Nason, 2008).

MBR: In this study, market-based CFP is measured by MBR. The original plan was to use Tobin's Q as a measure of market-based CFP. However, it is difficult to determine the replacement cost of a company's assets, which are needed for the calculation of Tobin's Q. Therefore, the MBR, the most used proxy for Tobin's Q in the literature, was used instead (Erickson & Whited, 2006). This measure has been considered by a large number of researchers in the existing literature to be an appropriate proxy for market-based CFP and has been used frequently (Lo & Sheu, 2007; Marti, Rovira-Val, & Drescher, 2015). Therefore, this measure was chosen for this study.

The MBR is calculated by dividing the market value by the book value of a firm's common stock. In order to determine a firm's market value, the number of shares outstanding must be multiplied by the share price, which equals the market capitalization of a firm (Laskar, Chakraborty, & Maji, 2017). To calculate a firm's book value, the total liabilities of a company must be subtracted from its total assets (Auret & Sinclaire, 2006).

3.3.2. Independent variables

The CSR activities of a firm constitute the basis of the independent variables in this work. Existing studies have used a wide variety of measures of CSR (Choi, Kwak, & Choe, 2010). Before measuring and quantifying CSR activity, these activities were first divided into CSR commitment and CSR communication. CSR commitment represents externally assessed and verified CSR activity, while CSR communication consists of CSR activities that have been reported by a company itself. The variables ESGScore, and SusAwards represent CSR commitment, while the variables CSRreport, CommuQuantity, and CommuBreadth represent CSR communication.

ESGScore: In this thesis, the Thomson Reuters' ESG Combined Score was used as the proxy for CSR commitment. Initially, it was anticipated to use the ESG scores constructed by one of the fourth largest sustainability rating agencies (MSCI ESG, ISS ESG, Video-Eiris, or Sustainalytics). While these scores are available in the Wharton Research Data Services database, accessing this information would have required a more extensive and expensive access than the one the university provides. After thorough literature research and consideration, the ESG Combined Score produced by Thomson Reuters was used as an alternative. The method behind the creation of this score is very similar to that published by ISS ESG, and produces comparable results (Doni & Johannsdottir, 2020). The Thomson Reuters ESG scores database was introduced in 2018. It replaces and extends the well-known ASSET4 database. The database contains time series data dating back to the year 2002, and entails ESG scores of more than 7000 international companies (Thomson Reuters, 2013).

A big weakness of sustainability scores like the ESG scores is that the company in question often provided most, if not all of the data used for the assessment, and that there is no kind of independent third-party verification of the accuracy

of the score (Orlitzky et al., 2003; Soana, 2011). ASSET4 and now Thomson Reuters ESG added an ethical screening to their method to overcome this issue. This screening includes any controversial information about a firm that has been published by NGOs, stock exchange filings, and news sources. The detection of this kind of information leads to the lowering of the overall score. Therefore, these scores are objectively measuring companies' ESG performance across the main areas, and differentiate firms that are transparent from those that are publicizing more than they are actually executing (Thomson Reuters, 2013). The three pillars, covering ten main areas, that are evaluated to estimate the ESG scores are the following: environmental aspects (resource use, emissions, innovation), social aspects (workforce, human rights, community, product responsibility) and governance aspects (management, shareholders, CSR strategy).

The sustainability performance of a firm in all ten areas is assessed. In a next step, this preliminary score is then offset against a controversial score, based on the negative information on the firm. This results in the ESG Combined Score. Thomson Reuters rates each company with a number between 0 and 1. The higher the score, the better the company's CSR performance. Based on this score, a grade is awarded to the firms. Investigating all of those aspects as well as taking into account controversies the companies were involved in, the Thomson Reuters ESG Combined Score offers a comprehensive rating that covers the multidimensional concept of CSR (Dinçer & Yüksel, 2018). To achieve the most accurate scoring, a combination of human and algorithmic sourcing is used to retrieve the data (Thomson Reuters, 2013).

Table 3 provides an overview of the Thomson Reuters ESG Combined Score range for every grade. Based on these grades, the firms were assigned a score for the variable. Consequently, the variable took a value between one and twelve.

SusAwards: Winning one or more sustainability awards was determined as another sign of CSR commitment. Previous studies (e.g., Klassen & McLaughlin, 1996) have used this variable in their research. Therefore, it was examined whether this variable also has an impact on the measures of CFP chosen in this study. In this study, not only environmental awards, but any kind of sustainability awards were taken into account. The information on whether a firm had been awarded with a sustainability prize was obtained from Thomson Reuters. The number of awards won was not taken into consideration. This variable was coded as a dummy variable. It therefore took the value of "1" if the company had won an award in the respective year, and "0" if not.

CSRreport: CSR reports are considered to be one of the most effective tools for a firm to inform a its stakeholders about its CSR commitments (Hooghiemstra, 2000). This variable was created to find out if stakeholders and shareholders actually value the publication of standalone CSR reports. Following Kolk and Perego (2010) and Gamerschlag, Möller, and Verbeeten (2011), this variable was operationalized as a dummy variable. It took the value "1" if a company released a standalone CSR report in the respective year, and "0" otherwise.

CommuQuantity: To understand whether the sheer quantity of CSR reporting has an impact on CFP, this variable was produced. In order to create this variable, it was necessary to create an overview of how voluminous the CSR reporting of the individual companies was in the in each year. It was decided to use the word count of the reports as an indicator. While this was more complex than using the number of pages, it proved to be the more accurate and appropriate metric, as the reports differed greatly in terms of font size and the use of images. For some companies, no standalone CSR reports were available, either for several years or generally. Some companies published socalled integrated reports in which the information on CSR was located in the same document as the financial information, but clearly separated from it. For some companies, CSR information was not explicitly presented and almost hidden in the annual reports. Assigning a word count of zero to companies that did not publish a standalone CSR report did not seem appropriate. Therefore, the annual and integrated reports of all firms, amounting in a total of 392 reports (49 firms over eight years), were screened manually for CSR-related information. The decision as to which text passages dealt with CSR-relevant topics was made by the author of this work herself. In order to get a picture of what is relevant to CSR, the SDGs were utilized. As the SDGs have the advantage of covering a wide range of sustainability issues, they were considered an appropriate orientation. The SDGs were studied in detail in advance so that the decision on what was considered as CSR-relevant could be made with confidence. The content related to CSR was then collected in separate word documents, and the number of words was determined. The word counts of the individual standalone CSR reports were also determined. These reports were not screened first because it could be assumed that the entire content is CSR related. To calculate the overall word count of the individual companies for the respective years, the word count of their annual or integrated report was added to the word count of the company's standalone CSR report, if available. On the basis of this word count, three groups were formed: high word count, medium word count, and low word count. The firms were then assigned to three groups: low communication quantity, medium communication quantity, and high communication quantity. The firms with an overall word count below 10500 were considered to be in the low communication quantity group, firms with a word count between 10500 and 25000 were assigned to the medium communication quantity group, and the firms with a word count above 25000 were selected into the high communication quantity group[4].

CommuBreadth: The role of the quantity of CSR reporting has been researched widely in the existing investment literature (Dhaliwal et al., 2012). A topic that has not received an equal amount of attention is the content of the reported information (Harjoto & Jo, 2015). A company that discloses CSR information across many dimensions of sustainability could

[4]The values 10500 and 25000 were about 2/3 and 4/3 of the mean of the word count, respectively, and where therefore chosen as limits

Table 3: ESG Combined Score range

Score Range	Grade	Variable score
0.0 <= score <= 0.083333	D-	1
0.083333 < score <= 0.166666	D	2
0.166666 < score <= 0.250000	D+	3
0.250000 < score <= 0.333333	C-	4
0.333333 < score <= 0.416666	C	5
0.416666 < score <= 0.500000	C+	6
0.500000 < score <= 0.583333	B-	7
0.583333 < score <= 0.666666	B	8
0.666666 < score <= 0.750000	B+	9
0.750000 < score <= 0.833333	A-	10
0.833333 < score <= 0.916666	A	11
0.916666 < score <= 1	A+	12

be considered to be a company engaged in many CSR policies and practices (Beck et al., 2018). For this study, a new coding system was developed that resulted in an index, which constitutes this variable.

The SDGs were used as a framework for the creation of this variable. These goals provide companies with a very broad and well-defined overview of sustainability issues. They transform the intangible topic of sustainability into tangible goals. No other sustainability agenda provides such detailed guidance on how to achieve CSR objectives. Therefore, their content was chosen to be the basis for the keywords used in the qualitative content analysis, which will be described in the following.

The SDGs were only introduced in October 2015, while the observation period of this paper spans from 2011 to 2018. This should not be a problem since, as already mentioned, only the thematic areas of the SDGs were addressed, which must have already been relevant before 2015. It has to be clarified that these goals specifically and the question of their legitimacy, were not the subject of this paper. They serve merely as an orientation. The single SDGs should therefore rather be seen as single CSR topics, and considered a helpful guide for navigating a company's CSR activities. Therefore, the single SDGs will be referred to as "CSR topics" in the following.

To come up with a catalogue of keywords, the progress reports on the SDGs of the last four years were screened manually[5]. These reports give an overview and an explanation of the CSR topics, and report on the measures taken to achieve each single CSR topic. During this screening, the words which were most frequently used by the UN authors in connection with each CSR topic were identified. The frequency of each word was noted. To avoid making the following content analysis too general, the number of keywords was limited to the 20 most frequently mentioned words per CSR topic.

A pretest was conducted to ensure that the keywords were reliably capturing the main message of the respective CSR topic. For this, 20 annual, integrated, and CSR reports were randomly selected from the sample, and uploaded to MAXQDA. In the next step, a code system for the keywords was created, resulting in 16 codes, one for each CSR topic[6,7]. The keywords were then entered into the lexical search of the program, and the coded sections in the reports were assigned to the corresponding CSR topic codes. After the screening, the highlighted passages were reviewed in order to see if the lexical search actually highlighted content that was related to the respective CSR topic. Based on these findings, modifications were made to the original set of keywords. The lexical search option in MAXQDA allows to set conditions for the keyword search. It is possible to make it a precondition that a keyword is only recognized as such if it is located in the same sentence as another predefined word. The other way around, it is possible to prohibit the program from finding a keyword in the text if it is mentioned in the same sentence as another predefined word. This allows to put keywords into context and prevents the highlighting of sections that have no relation to the information one is looking for. The main issue was that some words had a different meaning, depending on their context. For example, it had to be predefined that the word "rape" should not be highlighted if the words "seed" or "oil" were also in the same sentence. This was done to guarantee that the highlighted sections were addressing the issue of sexual abuse. Also, the word "AIDS", referring to the illness, had to be excluded from the list of keywords as it could not be satisfactorily distinguished from the word "aids", which refers to help or helping. Therefore, too many unrelated sentences would have been highlighted, and it was decided to exclude the word completely. Keywords with no hits in the pretest were also eliminated. This method of modification was repeated until the results reached a satisfactory level of accuracy. The list of keywords consisted of 150 words and

[5] These reports can be found in Appendix A4.

[6] Since Goal 17 of the SDGs is not a CSR topic, but an encouragement for the cooperation to achieve the other 16 goals, it was not included.

[7] A simplified graphic illustration of this process can be found in Appendix A2.

varied from five to 17 words per CSR topic (for an overview of this list, see Appendix A3).

In the next step, all annual and CSR reports[8] from the sample firms were screened with the code system in MAXQDA. The results were exported in an Excel document. The number of hits for the individual CSR topic in every single report was counted. If a company had issued both an annual and a standalone CSR report in one year, the hits in both reports for that year were added up. This resulted in an overview of the number of hits of every CSR topic for every firm in every year.

There were several ways in which this information could be processed and evaluated. As the number of assigned keywords was not the same for every single CSR topic, the idea to incorporate this irregularity was considered. In this scenario, the number of hits per topic would have been divided through the number of keywords assigned to this topic. This would have resulted in a hit count relative to the number of keywords. Another option was to conduct an absolute breadth score index. This index reflects how many CSR topics were addressed in a report in total. The latter option was considered to be more practical and appropriate for subsequent analysis and interpretation, and was, therefore, pursued. To conduct this index, it had to be determined from which number of hits a CSR topic was to be considered as having been addressed. If there were three or more hits assigned to a CSR topic, it was considered to have been captured in that year's reporting of a firm. This hurdle was set in order to guarantee that the keywords of a topic were not only found in a document due to an inaccuracy in the method. The hits were reviewed after each keyword search, but it could not be guaranteed that no mistake was ever overlooked. However, it was much more unlikely that three text passages in a document representing a false positive were overlooked during manual reviewing. The number of covered CSR topics was added up. The absolute breadth score index of a firm, therefore, took a value between 0 (no CSR topics were addressed in the reports of that year) and 16 (all CSR topics were addressed in the reports of that year). This index was used as an ordinally scaled variable in the regression.

3.3.3. Control variables

FirmSize: Size could have an influence on the value of CSR activities for a firm. Size influences the level of risk a firm faces. Firms that have a bigger presence in the market are exposed to more risk than smaller firms. With a greater firm size comes a greater amount of both internal and external transactions. This leads to a greater probability of negative events happening. Consequently, bigger firms should be more eager to engage in CSR activities than smaller firms to protect themselves from a loss of reputation in case of a negative outcome of one of their transactions. (Godfrey et al., 2009; Kimberly, 1976) The number of employees was used as a proxy for firm size. For each year and company, this value was extracted. The firms were assigned to three size groups. Firms with 1000 employees or less were assigned into the small group, firms with 1001 to 10000 employees were assigned to the medium group, and firms with more employees were assigned to the big group. Hence, the variable took a value of 1 to 3.

Industry: To avoid misleading interpretations, potential effects of the industry membership on CFP, triggered by, for example, fluctuations in the demand for goods or the reputation of certain industries over the years, had to be controlled for (Dess, Ireland, & Hitt, 1990). The industry membership might also influence the way CSR activities are perceived by stakeholders. Some industries might be suffering from a bad connotation (e.g. the tobacco industry) or a bad reputation because of past misconduct (Godfrey et al., 2009). CSR activities are used to build up moral capital by portraying the image of pursuing goals that are not only self-serving. Acting in a way that affects stakeholders in a negative way is counterproductive to those activities. Firms for industries that are generally known for having a negative social or environmental impact, CSR activities might be perceived to rather be some sort of 'blood money' than genuine efforts to improve the world. It might also be seen as an effort to substitute for destructive practices to offset the negative image of the firm. If this is the case, the value of the CSR activities either ceases or, in the worst case, those activities might be interpreted as an effort to deceive the public. (E. E. Jones, 1964) The allocation of the companies to the individual industries mostly followed the classification of Thomson Reuters Datastream. However, the tobacco industry was not considered to be part of the food industry, as it was perceived as much more controversial. Table 4 offers an overview of all industries.

FirmVisibility: There are a few ways in which firm visibility could influence CFP. The visibility of a firm has shown to influence its market value (Cheng et al., 2014). Also, as higher stakeholder pressures are usually placed on firms which are in the public eye, firm visibility needed to be controlled for (Shahzad & Sharfman, 2017). Increased public exposure and, therefore, higher visibility can put a firm under higher scrutiny by stakeholders, making it easier to expose any misconducts more easily than those of firms with less public exposure (K.-H. Kim, Kim, & Qian, 2018). Originally, marketing expenses were supposed to be a control variable for firm visibility. Unfortunately, the databases in the Wharton Research Data Services did only supply this information for a small number of the selected firms. Having to find a new set of firms that met all of the criteria, and searching for and scanning their annual and CSR reports would have taken up too much time. Consequently, after scanning the literature for appropriate and accessible alternatives, a proxy for firm visibility was generated. Taking a note from Gamerschlag et al. (2011), the visibility of a firm was determined by counting the number of hits when searching for the company's name on the Handelsblatt newspaper's archives website, while filtering for the respective year. This was considered an appropriate measure as the Handelsblatt is Germany's most important business-related newspaper. The fact that it is a German

[8] All 656 reports can be provided on request.

Table 4: Overview of industries

Number	Industry
1	Airlines
2	Financial Services
3	Real Estate
4	Energy
5	Food Processing & Retailing
6	Integrated Telecommunications Services
7	Insurance
8	Oil & Gas
9	Personal Products
10	Pharma
11	Tobacco

newspaper was not considered to distort the results as all the firms were listed companies and of international interest. To be sure, the hit counts of firms from Germany and other countries which were relatively similar in size and popularity were compared, and no bias towards German companies was detected. The firms were subsequently sorted into three groups. Firms with 25 or viewer hits were assigned to the low visibility group, firms with a hit count between 26 and 150 were assigned to the medium visibility group, and firms with a higher hit count were assigned to the high visibility group. Accordingly, the variable took a value between 1 and 3.

FirmAge: The age of a firm can also play a role in the financial performance of a company. Many studies assume that, as the years go by, only the better companies will survive, and therefore it can be assumed that older companies will continue to have a higher life expectancy (Dunne, Roberts, & Samuelson, 1989). It has also been shown that older companies enjoy higher profits and value (Hopenhayn, 1992). Therefore, the study controlled for firm age.

MAacquirer & MAtarget: It was also controlled for whether the firms had been the target or an acquirer in a merger or an acquisition with a volume of over 1 million dollars. This amount was chosen, as the firms in the sample were in the 80th percentile with regard to market capitalization. So smaller volume M&A cases seemed neglectable. Because a differentiation between being a target or being an acquirer was made, two dummy variables were created.

The table 5 offers an overview of all variables.

3.4. Statistical method

In the following section, the choice of the regression models is briefly explained. An overview of all hypotheses, and the models in which they are being tested, is presented. All of the analyses were conducted in the statistical software Stata/IC Version 16.1.

3.4.1. Panel data

A great advantage of panel data is that individual heterogeneity can be controlled for. Not controlling for this heterogeneity can result in biased results (Moulton, 1987). Panel data, compared to, for example, time-series studies, give more informative data and less collinearity among the variables, since the cross-section dimension adds variability. This more informative data also produces more reliable results. (Baltagi, 2005)

3.4.2. Fixed-effects and random-effects models

The two most commonly used techniques to analyze panel data are fixed-effects and random-effects models. Both models have their advantages and disadvantages. Due to the within-transformation, the fixed-effects model has the advantage of being able to control individual, unnoticed effects and is, therefore, more robust against the assumption of strict exogeneity. However, this within-transformation also contains a disadvantage. In the course of the transformation, all, also observable, time-invariant effects are removed, so that constant, independent variables cannot be considered (Wooldridge, 2005). Unlike the fixed-effects model, the random-effects model assumes that the individual, unobserved effects are not correlated with the exogenous variables. The unobserved effect has an expected value from zero and is, together with the idiosyncratic error, part of the disturbance term (Baltagi, 2005). While the random-effects model takes time invariant effects into account, and often produces a more generalizable result, it often introduces biases in the estimates of ß (Baltagi, Bresson, & Pirotte, 2003).

3.4.3. Hausman test

The question of whether to use the fixed-effects or random-effects model when analyzing panel data has been debated thoroughly, with advocates for both models (Wooldridge, 2005). In order to determine whether the random-effect model or the fixed-effects model is more appropriate, a Hausman test was conducted (Greene, 2002). The Hausman test is a specification test that is based on the difference between the random- and the fixed-effects estimator (Wooldridge, 2005). This test reveals which of the two tests is most efficient given the sample. It uses a χ^2 distribution with one degree of freedom (Greene, 2002)

Table 5: Overview of variables

Dependent variables	
ROA	Return-on-assets
MBR	Market-to-book ratio
Independent variables	
ESGScore	Thomson Reuters ESG Combined Score
SusAwards	Was a sustainability award won?
CSRreport	Was a standalone CSR report published?
CommuQuantity	Word count of CSR reporting, 3 groups
CommuBreadth	Number of CSR topics communicated
Control variables	
FirmSize	Number of employees, 3 groups
Industry	Industry membership
FirmVisibility	Articles in Handelsblatt archive, 3 groups
FirmAge	Age of the firm in years
MAacquirer	Was the firm the acquirer in an M&A deal?
MAtarget	Was the firm the target in an M&A deal?

and tests the null hypothesis that the conditional mean of the disturbances given the regressor is zero (Baltagi et al., 2003). If the chi2 takes a value smaller than 0.05, the null hypothesis is rejected (Greene, 2002), and the fixed-effects estimator should be reported. If the null hypothesis cannot be rejected, the random-effects estimator is presented (Baltagi et al., 2003).

3.4.4. Mixed-effects models

According to the literature, multilevel mixed-effects models have been used to test panel data as well (e.g., Pablo-Romero & Sánchez-Braza, 2017). Mixed-effects models are regression models that fixed as well as contain group-specific random effects (Speelman, Heylen, & Geeraerts, 2018). In longitudinal data and panel data, these random effects prove to be useful for modeling intracluster correlation (StataCorp, 2013). Mixed-effects models can yield more generalizable results and allow for time-invariant effects, like industry, to be considered in the regression. These effects are omitted in a fixed-effects model. (Speelman et al., 2018) Therefore, to ensure the robustness of the results of this study and measure the influence of time-invariant variables, a multilevel mixed-effects regression was performed for every model and the results were compared to those of the random- and fixed-effects regressions. The table 6 offers an overview of which hypotheses are tested in which models.

4. Results

In the following chapter, the results of the statistical analysis will be presented. The section is divided into two parts, starting with descriptive statistics which already contains a high information content. This is followed by the results of the regression analyses. These results determine whether the hypotheses can be supported or must be rejected.

4.1. Descriptive statistics

Sample description: The data set used in this analysis was a fixed, strongly balanced panel, since the same set of firms was examined for each period, and no values for any variables are missing in any period (Greene, 2002). This was an excellent precondition for the regression analysis (Baltagi, 2005). The sample consisted of 49 firms from 14 European countries. The observation period was eight years. The number of observations was, therefore, 392. The firms in the sample, on average, were 74 years old, employed 68867 people, and had a market capitalization of 75.400.000 US Dollars. Table 7 offers an overview of the distribution of companies by industry.

ESG Scores: On average, the firms in the sample had the highest ESG Scores in 2015 (7.5) and the lowest in 2013 and 2014 (6.4). The industries whose firms had the highest ESG Scores on average were the real estate (7.93), energy (7.69), and insurance (7.42) industry. The industries with the lowest average ESG scores were the airlines (5.69), tobacco (6.13), and financial services (6.24) industry. The average ESG Score over all industries and years was 6.73. The firms in the medium size group had the highest ESG score on average (7.14).

Sustainability awards: In the year 2011, the highest number of firms (about 74%) won one or more sustainability awards. In the following years, the percentage was between 63 and 65.

The industries that have won one or more sustainability awards in the most years are the integrated telecommunications services, personal products, and tobacco (all 100%) industry. The industries that won such awards in the lowest number of years were the airlines (43.75%), financial services (48.61%), and oil and gas (50%) industry.

Small firms with regard to employees won sustainability awards most frequently, while firms in the midsize category won them in the lowest number of years.

Table 6: Hypotheses and regression models

Hypotheses	Model
H_{1a}: Being awarded a high CSR score has a positive influence on accounting-based CFP. H_{1b}: Being awarded a sustainability award has a positive influence on accounting-based CFP.	Random-effects and mixed-effects ROAi, t = ß0 + ß1 * Awardi,t + ß2 * ESGScorei,t + ß3 * FirmSizei,t + ß4 * FirmVisibilityi,t + ß5 * FirmAgei,t + ß6 * MAtargeti,t + ß7 * MAacquireri,t + ß8 * Industryi + $\varepsilon_{i,t}$
H_{2a}: Being awarded a high CSR score has a positive influence on market-based CFP. H_{2b}: Being awarded a sustainability award has a positive influence on market-based CFP.	Random-effects and mixed-effects MBRi, t = ß0 + ß1 * ESGScorei,t + ß2 * Awardi,t + ß3 * FirmSizei,t + ß4 * FirmVisibilityi,t + ß5 * FirmAgei,t + ß6 * MAtargeti,t + ß7 * MAacquireri,t + ß8 * Industryi + $\varepsilon_{i,t}$
H_{3a}: Issuing a standalone CSR report has a positive influence on accounting-based CFP. H_{3b}: A higher quantity of CSR reporting has a positive influence on accounting-based CFP. H_{3c}: Addressing a broader range of CSR topics has a positive influence on accounting-based CFP.	Fixed-effects and mixed-effects ROAi, t = ß0 + ß1 * SusReporti,t + ß2 * CommuQuantityi,t + ß3 * CommuBreadthi,t + ß4 * FirmSizei,t + ß5 * FirmVisibilityi,t + ß6 * FirmAgei,t + ß7 * MAtargeti,t + ß8 * MAacquireri,t + ß9 * Industryi + $\varepsilon_{i,t}$
H_{4a}: Issuing a standalone CSR report has a positive influence on market-based CFP. H_{4b}: A higher quantity of CSR reporting has a positive influence on market-based CFP. H_{4c}: Addressing a broader range of CSR topics has a positive influence on market-based CFP.	Fixed-effects and mixed-effects MBRi, t = ß0 + ß1 * SusReporti,t + ß2 * CommuQuantityi,t + ß3 * CommuBreadthi,t + ß4 * FirmSizei,t + ß5 * FirmVisibilityi,t + ß6 * FirmAgei,t + ß7 * MAtargeti,t + ß8 * MAacquireri,t + ß9 * Industryi + $\varepsilon_{i,t}$

Table 7: Number of firms per industry

Number	Industry	Firms
1	Airlines	4
2	Financial Services	9
3	Real Estate	5
4	Energy	4
5	Food Processing & Retailing	9
6	Integrated Telecommunications Services	1
7	Insurance	3
8	Oil & Gas	5
9	Personal Products	2
10	Pharma	6
11	Tobacco	1

Standalone CSR reports: The most standalone CSR reports were published in 2012, 2013 and 2015 (in all years 71.4%). The year in which the least standalone reports were published was 2011 (53%). Over all industries and years, 67% of firms published standalone CSR reports. The industries which tended to publish more standalone CSR reports were the tobacco (100%), oil and gas (97.5%), and insurance (95.8%) industry. Least likely to publish standalone CSR reports were the firms in the pharma (27%), food processing and retailing (47%), and real estate (50%) industry. The firms with the highest number of employees published standalone CSR reports most frequently.

Communication quantity: There was no consistent increase in the quantity of reporting over the observation period. Looking at the absolute word count of CSR reporting, there was even a decrease in 2018. The industries whose companies most often belonged to the highest communication quantity group were the oil and gas, integrated communications services and energy industry. The industries with the lowest levels of communication quantity were the food processing and retailing, pharma and real estate industry. Firms with a higher number of employees also communicated more in terms of quantity.

Communication breadth: There was an increase in the

overall CSR breadth from 2011 (ø5.12 CSR topics) to 2018 (ø7.10 CSR topics), with a drop in 2016.

Many CSR topics were communicated by the highest percentage of firms in 2018 and the lowest percentage of firms in 2011. Some topics experienced peaks and lows in other years. The topic of "no poverty" was communicated the most in 2015 and 2016, and the least in 2011, 2012 and 2013. The topic of "good health and well-being" peaked in 2013 and was at a lowest in 2017. The communication of almost all the CSR topics increased sharply from 2011 to 2012, except the topics of "zero hunger" and "reduced inequalities", which actually slightly decreased in 2012. Also, the communication of many goals experienced a sharp decrease in the year 2016. Exceptions for this were the topics of "no poverty", "zero hunger", and "clean water and sanitation", which stayed the same compared to 2015, and the topic of "sustainable cities and communities", which slightly increased. 2016 was actually the year in which the topic of "no poverty" was communicated the most. The CSR topics with the sharpest decreases in 2016 were "good health and well-being (19% decrease), "quality education" (10% decrease), "industry, innovation, and infrastructure" (10% decrease) and "life on land" (16% decrease).

The CSR topics that were communicated the most were "peace, justice, and strong institutions" (85.71%), "affordable and green energy" (77.55%), "climate action" (74.75%), and "responsible consumption and production" (62.25%). The CSR topics that were communicated the least were "life below water" (2.3%), "reduced inequalities" (2.3%), "no poverty (9,44%), and "life on land" (15.56%).

The industries that covered the most CSR topics in their reporting were the energy (ø8.09 CSR topics), pharma (ø7.75 CSR topics) and insurance industry (ø7.25 CSR topics). The industries that covered the least CSR topics were the tobacco (ø4.13 CSR topics), real estate (ø4.7 CSR topics), and the airline industry (ø5.31 CSR topics).

Small firms reported on 4.7, midsize firms on 5.2, and big firms on 6.6 CSR topics on average.

4.2. Regression results

In the following, the regression results will be presented. In this course, the hypotheses put forward in chapter 2 are tested and subsequently either supported or rejected. A 95% confidence interval was chosen throughout this study. Therefore, the significance level is equal to 5%. Consequently, pvalues which are equal to or smaller than 0.05 will be interpreted as statistically significant with a 95% accuracy.

CSR commitment on ROA: The pvalue produced by the Hausman test (Prob > chi2 = 0.1440), was above the critical value of 0.05. This indicates that the random-effects model gives consistent estimates for the model measuring the influence of CSR engagement on ROA. Therefore, the random-effects model was a more appropriate model in this case, and used in the regression analysis.

Table 8 shows the output of the random-effects GLS regression examining the influence of CSR commitment on ROA. The overall R^2 of the model was 0.4192. Therefore, the model could explain 41.92% of the variance of ROA. The effect of the ESGScore variable on ROA was statistically not significant. With a pvalue of 0.279, there is no evidence that the Thomson Reuters ESG Combined Scores have any effect on a firm's ROA. The results reveal that the effect of the SusAwards variable had no significant on ROA either (p=0.609). The effects of the control variables FirmVisibility (Group 2, p=0.424; Group 3, p=0.577), FirmAge (p=0.966), FirmSize (Group 2, p=0.869; Group 3, p=0.678), MAacquirer (p=0.519) and MAtarget (p=0.826) were also not significant. Only the membership in industry 10 (ß=11.47778, p=0.000) and industry 11 (ß=14.24037, p=0.013) had a positive and significant effect on ROA.

Table 9 shows the results of the mixed-effects ML regression examining the influence of CSR commitment on ROA. The direction of the effects did not change, and the size of the effects were similar to those of the random-effects model. Both the effect ESGScore and the SusAwards variable were also negative and insignificant in this model. Industry 10 still had a positive and significant effect (ß=5.911365, p=0.008) on ROA. In the mixed-effects model, industry 11 still had a positive but no longer significant effect on ROA. Industry 2 on the other hand was negative and significant (ß=-4.637318, p=0.030) in this model.

Based on these results, the hypothesis, H_{1a}, that having a higher ESG Combined Score has a positive influence on accounting-based CFP must be rejected. Also, the hypothesis, H_{1b}, that being awarded one or multiple sustainability awards has a positive influence on accounting-based CFP must be rejected.

CSR commitment on MBR: The p-value produced by the Hausman test (Prob > chi2 = 0.0802), was above the critical value of 0.05. Therefore, the random-effects model was chosen as the more appropriate model in this case.

Table 10 shows the output of the random-effects GLS regression examining the influence of CSR commitment on MBR. The overall R^2 of the model was 0.5111. Therefore, the model could explain 51.11% of the variance of the dependent variable. The regression shows that the effect of the ESGScore variable on MBR was statistically not significant (p=0.290), and neither was the effect of the SusAwards variable (p=0.104). The effects of the control variables FirmVisibility (Group 2, p=0.776; Group 3, p=0.795), FirmAge (p=0.385), FirmSize (Group 2, p=0.911; Group 3, p=0.933), MAacquirer (p=0.255), and MAtarget (p=0.750) on MBR were also not significant. Again, from all industries, only the membership in industry 10 (ß=6.253643, p=0.000), and industry 11 (ß=6.18611, p=0.017) had a strong and positive and significant effect on MBR.

Table 11 shows the results of the mixed-effects ML regression examining the influence of CSR commitment on MBR. The results of this model support the results on the size and direction of the effects produced by the random-effects model. Both the ESGScore and the SusAwards variable were still negative and not significant. Industry 10 still had a positive and significant effect (ß=6.261066, p=0.000) on ROA.

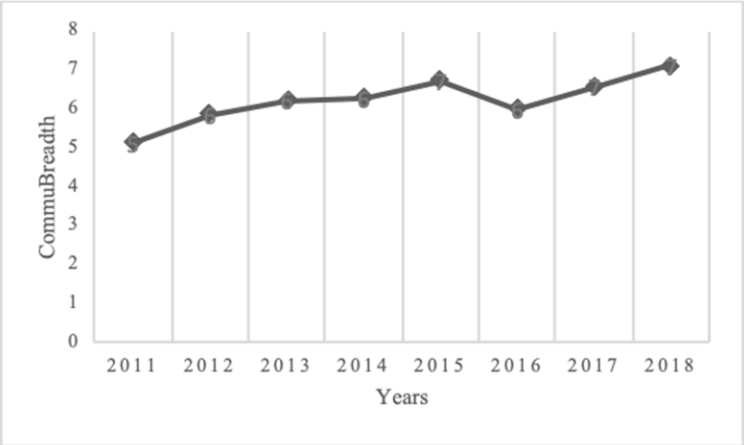


Figure 2: Communication breadth over the years

Table 8: CSR commitment on ROA: Random-effects regression results

	Coefficient	z	P>|z|
ESGScore	-.1521758	-1.08	0.279
SusAwards	-.2941941	-0.51	0.609
FirmVisibility			
2	.6561508	0.80	0.424
3	.6190387	0.56	0.577
FirmAge	-.0005289	-0.04	0.966
MAacquirer	.3918068	0.65	0.519
MAtarget	-.1359419	-0.22	0.826
FirmSize			
2	.3822286	0.16	0.869
3	-1.053903	-0.42	0.678
Industry			
2	-3.141814	-1.02	0.308
3	3.495334	0.86	0.388
4	2.976599	0.80	0.422
5	3.242277	1.05	0.296
6	-1.976519	-0.35	0.729
7	-2.367294	-0.60	0.546
8	1.797051	0.53	0.598
9	8.207561	1.86	0.063
10	11.47778	3.48	0.000
11	14.24037	2.48	0.013
Dependent variable: ROA			
R^2=.4192			
Number of obs.=392			

In the mixed-effects model, Industry 11 still had a positive and significant effect on ROA. In contrast with the results of the random-effects model[9], the membership in industry 9 also had positive and significant effect on MBR (ß=3.8328, p=0.019).

The results reveal that having a higher ESG Combined Score does not have a significant positive influence on CFP. Therefore, H_{2a} has to be rejected. Also, as being awarded one or multiple sustainability awards did not show to have a significant positive influence on CFP, H_{2b} has to be rejected as well.

CSR communication on ROA: The p-value indicated by the Hausman test (Prob > chi2 = 0.0371), was below the critical value of 0.05. This indicates that the null hypothesis for the Hausman test can be rejected and that the fixed-effects model

[9]The p-value of industry 9 was 0.065 in the random-effects model, and, therefore, not far from being significant.

Table 9: CSR commitment on ROA: Mixed-effects regression results

	Coefficient	z	P>|z|
ESGScore	-.1608864	-1.20	0.231
SusAwards	-.324592	-0.59	0.556
Industry			
2	-4.637318	-2.17	0.030
10	5.911365	2.67	0.008
11	11.47489	1.56	0.118

Dependent variable: ROA
Number of obs.=392

Table 10: CSR commitment on MBR: Random-effects regression results

	Coefficient	z	P>|z|
ESGScore	.051698	1.06	0.290
SusAwards	-.3265301	-1.63	0.104
FirmVisibility			
2	.0821697	0.28	0.776
3	.1023839	0.26	0.795
FirmAge	.004871	0.87	0.385
MAacquirer	-.238557	-1.14	0.255
MAtarget	.0681623	0.32	0.750
FirmSize			
2	.0938445	0.11	0.911
3	-.0777448	-0.08	0.933
Industry			
2	-.5558501	-0.40	0.691
3	-.6338108	-0.37	0.715
4	.7152326	0.43	0.667
5	1.29386	0.93	0.354
6	-.4029822	-0.16	0.876
7	-.6371989	-0.36	0.720
8	.0842268	0.05	0.956
9	3.687619	1.84	0.065
10	6.253643	4.20	0.000
11	6.18611	2.39	0.017

Dependent variable: MBR
R^2=.5111
Number of obs.=392

Table 11: CSR commitment on MBR: Mixed-effects regression results

	Coefficient	z	P>|z|
ESGScore	0.0525767	-1.09	0.277
SusAwards	-.3306733	-1.67	0.095
Industry			
9	3.8328	2.34	0.019
10	6.261066	5.11	0.000
11	6.397031	3.02	0.003

Dependent variable: MBR
Number of obs.=392

gives consistent estimates for the model measuring the influence of CSR communication on ROA. Therefore, the fixed-effects model, being the more appropriate model in this case, was chosen for the regression analysis.

Table 12 shows the output of the fixed-effects (within) regression examining the influence of CSR communication on ROA. The overall R^2 of the model was 0.0007. Therefore, the model could explain only 0.07% of the variance of ROA. The influence of the CSRreport variable on ROA was positive, but not statistically significant (p=0.247). A high score in CommuQuantity showed a negative and statistically significant influence on ROA (ß=-2.315213, p=0.032). The effect of the variable representing communication breadth was slightly positive but statistically not significant (p=0.559). The effect of the FirmVisibility variable on ROA was positive but also not significant (Group 2, p=0.140; Group 3, p=0.095). The results of this model suggest, that firm age had a negative and statistically significant influence on ROA (ß=-.2558694, p=0.003). MAacquirer had a positive, and MAtarget a negative influence on ROA, but both effects were statistically not significant. The FirmSize variable did not have a significant influence on ROA either.

Table 13 shows the results of the mixed-effects ML regression examining the influence of CSR communication on ROA. The results were slightly different to those of the fixed-effects model. The effect of the CSRreport variable was still positive (ß=2.000287) but now statistically significant (p=0.024). The effect of belonging to the highest scoring group of the CommuQuantity variable was still negative and statistically significant (ß=-2.531776, p=0.012). In contrast to the fixed-effects model, the mixed-effects model also considered time invariant variables like Industry. Industry 10 showed a positive and statistically significant (ß=5.117817, p=0.010) influence on ROA. The membership in both industry 2 (ß=-5.098802, p=0.008) and industry 7 (ß=-5.120999, p=0.036) had positive and statistically significant in-fluence on ROA. The effect of communication breadth on ROA was now slightly negative, but still not statistically significant.

Based on the results, the hypotheses were evaluated. While the influence of issuing a standalone CSR report on accounting-based CFP was positive but not statistically significant in the fixed-effects model, the mixed-effects model showed a positive and statistically significant effect. H_{3a} can, therefore, only be supported with some reservation. A higher quantity of reporting did not, as assumed, have a positive, but a significant and negative effect on ROA. Therefore, H_{3b} has to be rejected. The results also show that addressing a broader range of sustainability issues did not have a significant and positive influence on accounting-based CFP. Consequently, H_{3c} has to be rejected, too.

CSR communication on MBR: The p-value indicated by the Hausman test (Prob > chi2 = 0.0274), was below the critical value of 0.05. Therefore, the fixed-effects model was chosen in this analysis.

Table 14 shows the output of the fixed-effects (within) regression examining the influence of CSR communication on MBR. The overall R^2 of the model is 0.0012. Therefore, the model could explain only 0.12% of the variance of MBR. The influence of the variable CSRreport was positive, but not statistically significant (p=0.423). Being part of the highest scoring group with regard to communication quantity had a negative and significant but not statistically significant influence on MBR (ß=-.6105019, p=0.10). The effect of the CommuBreadth variable was positive and statistically significant (ß=0.1185186, p=0.011). The effect of the control variables FirmVisibility (Group 2, p=0.654; Group 3, p=0.848), FirmSize (Group 2, p=0.738; Group 3, p=0.526), MAacquirer (p=0.334), and MAtarget (p=0.755) were all statistically not significant. The results indicate that FirmAge had a positive and significant influence on MBR (ß=.0682975, p=0.021).[10]

Table 15 shows the results of the mixed-effects ML regression examining the influence of CSR communication on MBR. The direction of the effects did not change, and the size of the effects are similar to the results of the fixed-effects model. The effect of the CSRreport variable is still positive and not statistically significant (p=0.468). The direction of the influence of high CommuQuantity was still negative (ß=-0.843428) and now statistically significant (p=0.018). The influence of CommuBreadth was still positive and statistically not significant (ß=-.1419855, p=0.001). Unlike the fixed-effects model, the mixed-effects model also considers time invariant variables like industry membership. Memberships in industry 9 (ß=3.322546, p=0.043), industry 10 (ß=5.701553, p=0.000) and industry 11 (ß=6.149283, p=0.004) all showed a positive and statistically significant influence MBR.

These results show that issuing a standalone CSR report does not have a statistically significant positive influence on market-based CFP. Therefore, H_{4a} has to be rejected. Also, a higher quantity of CSR reporting is shown to have a negative influence on market-based CFP. Consequently, H_{4b}, has to be rejected as well. As expected, addressing a broader range of sustainability issues has a positive and significant influence on market-based CFP. H_{4c} can therefore be supported. Table 16 provides a final overview of all hypotheses and whether they were supported or rejected.

5. Follow-up studies

During the implementation of the main study, it became apparent that the data that was collected and processed was far too rich to end the data analysis by simply answering the hypotheses. Therefore, followup studies were conducted.

5.1. Followup study 1

5.1.1. Purpose and method

Not only the question of whether CSR efforts enhance or harm the financial performance of firms is much debated, but

[10] As firm age only showed a significant influence on CFP when the industry variable was omitted from the model, it can be assumed that the age of a firm has no influence on both accounting-based and market-based CFP.

Table 12: CSR communication on ROA: Fixed-effects regression results

	Coefficient	t	P>|t|
CSRreport	1.172592	1.16	0.247
CommuQuantity			
2	-.8275635	-0.94	0.349
3	-2.315213	-2.16	0.032
CommuBreadth	.0792509	0.59	0.559
FirmVisibility			
2	1.258785	1.48	0.140
3	2.025527	1.67	0.095
FirmAge	-.2558694	-3.00	0.003
MAacquirer	.2290191	0.38	0.704
MAtarget	-.168377	-0.27	0.784
FirmSize			
2	1.245724	0.47	0.637
3	-.1002454	-0.03	0.973

Dependent variable: ROA
R^2=.0007
Number of obs.=392

Table 13: CSR communication on ROA: Mixed-effects regression results

	Coefficient	z	P>|z|
CSRreport	2.000287	2.25	0.024
CommuQuantity			
2	-1.132748	-1.34	0.179
3	-2.531776	-2.50	0.012
CommuBreadth	-.1110395	-0.93	0.354
Industry			
2	-5.098802	-2.67	0.008
7	-5.120999	-2.09	0.036
10	5.117817	2.57	0.010

Dependent variable: ROA
Number of obs.=392

there is also discussion around the question if the communication of all types of CSR topics has the same effect or if the individuals topics have a different effect on reputation (Choi et al., 2010; Rodriguez-Fernandez, 2016). The focus of the first followup study was the question of whether the communication of specific goals has different influences on CFP. In order find out, the variable for CSR communication breadth was replaced with dummy variables for the single CSR topic, taking the value of "1" if the CSR topic was reported, and "0" if not.

5.1.2. Results

CSR topics on ROA: The results of both the fixed-effects[11] and mixed-effects regression showed that none of the individual CSR topics proved to have a statistically significant influence on ROA. The results can be found in the Appendix A4.

CSR topics on MBR: The p-value indicated by the Hausman test (Prob > chi2 = 0.5743), was above the critical value of 0.05. Therefore, the random-effects model was chosen as it was the appropriate model in this case.

The results of the random-effects regression (see table 17) show that the communication of three CSR topics (topic 5, topic 7 and topic 10) had a positive and statistically significant influence on MBR, with topic 10 having the greatest influence (ß=2.453523, p=0.000). The communication of topic 9 had a negative and statistically significant influence on MBR. The influence of high communication quantity is still negative and statistically significant (ß=-.7519658, p=0.032).

The results of the mixed-effects regression (see table 18) also show that topic 5, topic 7 and topic 10 had a positive and statistically significant influence on MBR, while the influence

[11] The p-value indicated by the Hausman test (Prob > chi2 = 0.0005), was below the critical value of 0.05.

Table 14: CSR communication on MBR: Fixed-effects regression results

	Coefficient	t	P>|t|
CSRreport	.2790973	0.80	0.423
CommuQuantity			
2	-.1204292	-0.40	0.692
3	-.6105019	-1.65	0.100
CommuBreadth	.1185186	2.54	0.011
FirmVisibility			
2	.1314942	0.45	0.654
3	.0796989	0.19	0.848
FirmAge	.0682975	2.32	0.021
MAacquirer	-.200742	-0.97	0.334
MAtarget	.066247	0.31	0.755
FirmSize			
2	.3039509	-0.33	0.738
3	.6374929	-0.63	0.526

Dependent variable: MBR
R^2=.0012
Number of obs.=392

Table 15: CSR communication on MBR: Mixed-effects regression results

	Coefficient	z	P>|z|
CSRreport	.2424482	0.73	0.468
CommuQuantity			
2	-.2509538	-0.85	0.398
3	-.843428	-2.36	0.018
CommuBreadth	-.1419855	3.29	0.001
Industry			
9	3.322546	2.02	0.043
10	5.701553	4.59	0.000
11	6.149283	2.89	0.004

Dependent variable: MBR
Number of obs.=392

Table 16: Overview of hypotheses and conclusions

CSR commitment	
H_{1a}: Being awarded a high CSR score has a positive influence on accounting-based CFP.	×
H_{1b}: Being awarded a sustainability award has a positive influence on accounting-based CFP.	×
H_{2a}: Being awarded a high CSR score has a positive influence on market-based CFP.	×
H_{2b}: Being awarded a sustainability award has a positive influence on market-based CFP.	×
CSR disclosure	
H_{3a}: Issuing a standalone CSR report has a positive influence on accounting-based CFP.	∼
H_{3b}: A higher quantity of CSR reporting has a positive influence on accounting-based CFP.	×
H_{3c}: Addressing a broader range of CSR topics has a positive influence on accounting-based CFP.	×
H_{4a}: Issuing a standalone CSR report has a positive influence on market-based CFP.	×
H_{4b}: A higher quantity of CSR reporting has a positive influence on market-based CFP.	×
H_{4c}: Addressing a broader range of CSR topics has a positive influence on market-based CFP.	✓

Table 17: Single CSR topics on MBR: Random-effects regression results

	Coefficient	z	P>|z|
CSRreport	.497708	1.46	0.145
CommuQuantity			
2	.0837933	0.27	0.787
3	-.7519658	-2.15	0.032
Topic5	.5544662	2.99	0.003
Topic7	.6288552	2.50	0.013
Topic9	-.6080846	-2.92	0.004
Topic10	2.453523	5.43	0.000
Dependent variable: MBR			
R^2=.5770			
Number of obs.=392			

Table 18: Single CSR topics on MBR: Mixed-effect regression results

	Coefficient	z	P>|z|
CSRreport	.4575111	1.42	0.155
CommuQuantity			
2	-.0592197	-0.20	0.842
3	-.8920096	-2.66	0.008
Topic5	.5779647	3.26	0.001
Topic7	.5952299	2.49	0.013
Topic9	-.5990344	-3.04	0.002
Topic10	2.602657	5.96	0.000
Dependent variable: MBR			
Number of obs.=392			

of topic 9 remained negative and statistically significant.

5.2. Followup study 2

5.2.1. Purpose and method

As explained in chapter 1, firms may display a communication-behavior gap when it comes to their CSR reporting. The question of whether this strategy is profitable for a firm or punished by its stakeholders is still under debate. The purpose of the following procedure was to disclose whether there was indeed a difference in the level of communication and commitment for certain companies in the sample. While the variables representing CSR commitment had mostly negative (albeit insignificant) influence on both accounting-based and market-based CFP, the variables representing CSR communication had positive and negative influences on the CFP measures. These results did not directly indicate whether a company could benefit from communicating a lot while showing little real commitment. Based on the findings from the literature presented in chapter 1, it is assumed in this study that many companies in the sample exhibit such a difference. The question of whether stakeholders and shareholders can recognize such a discrepancy and punish it accordingly is somewhat more complex. It can be expected that shareholders will become more familiar with the companies in which they eventually invest money. It is therefore assumed that a high level of CSR communication with a low level of CSR commitment has a negative impact on market-based CFP. According to this line of reasoning, this behavior should have a positive impact on accounting-based CFP.

To quantify the firms' behavior, four groups, to which all companies in the sample were assigned annually, were created (see table 19).

Before labeling a firm as "high/low level CSR commitment" or "high/low level CSR communication", it had to be specified what should be considered as a high score in the independent variables. If a firm had an ESG Combined score of seven or higher, it was considered to be in the better half of the sample[12]. If a firm had won at least one sustainability award in the respective year (so scored "1" for the awards dummy) and was in the better half regarding ESG scores, it was considered to display a high level of CSR commitment. Belonging to the highest reporting quantity group (group 3), having reported on 7 or more CSR topics (the median of the CommuBreadth variable was 6), and the publishing of a standalone CSR report (scoring "1" for the CSRreport dummy) were all considered to be signs of a high level of CSR communication. If a firm showed two out of these three signs,

[12]The number seven was chosen because the median of the ESG score variable was 6, so the next higher number was considered to be above average.

Table 19: Groups based on commitment and communication levels

Group 1	Low commitment, low communication
Group 2	Low commitment, high communication
Group 3	High commitment, low communication
Group 4	High commitment, high communication

it was considered to display a high level of CSR communication. According to these assessments, the firms were assigned to the four groups. This was done for every firm and every year. These assignments constituted a new variable, “CombGroup”, which took a value between one and four. The regression models were chosen in the same way as described in the above.

In a next step it was examined whether the general affiliation with "high level commitment" or "high level communication" has a significant influence on CFP. To test this, dummy variables were created for these two groups. These variables were labelled “HighCommit” and “HighCommu”. This was done to see whether the effect of the four groups were actually due to the constellation of their levels of commitment and communication.

5.2.2. Results

Descriptive statistics show that the largest percentage of enterprises over the entire observation period (41.1%) were in the second group. Although the percentages vary over the years, no linear increase is apparent. The second largest group in percentage terms was that of companies that display low levels of both CSR commitment and CSR communication. The group of companies that show a high level of commitment, but a low level of communication was the group with the lowest percentage.

Group membership on ROA: Neither the results of the random-effects regression[13], nor those of the mixed-effects regression showed a significant effect of any group on ROA. Also, for high level commitment, the results of both the random-effects[14] and the mixed-effects regression showed a negative, but statistically not significant effect on ROA. The same was the case for high level communication.

Group membership on MBR: After performing and interpreting the Hausman test, a random-effects and a mixed-effects regression were performed to determine the influence of group membership on MBR.

The random-effects model showed no statistically significant effect of any of the individual groups on MBR. However, the effect of group 2 (ß=-.3917638, p=0.068) was close to being statistically significant (see table 20) .

The results of the mixed-effects regression (see table 21) showed that being a part of group 2 had a negative and statistically significant influence on MBR (ß=-0.4167528, p=0.050).

The results of the random-effects and mixed-effects regressions examining the influence of high-level commitment on MBR showed that demonstrating a high level of commitment had a positive but statistically insignificant impact on MBR. The results of the fixed-effects and mixed-effects regressions examining the influence of high-level communication showed that demonstrating a high level of communication had a negative, but also statistically insignificant impact on MBR[15].

6. Discussion

Firms face increased pressure from their stakeholder and shareholders to increase their commitment to CSR and to transparently report on this commitment. While these activities are costly and time-consuming, they have been associated with an improved CFP. This relationship between CSR and CFP has opened up a new research field, which has been extensively explored over the past decades. However, most of the studies examining this relationship do not distinguish between CSR commitment and CSR communication in their assessment of CSR (Orlitzky et al., 2003). In the present study, this shortcoming is addressed and rectified. This thesis further contributes to the strand of research that critically addresses potential discrepancies between the level of CSR commitment and CSR communication of firms.

In this chapter, the descriptive and regression results of the main study are discussed first. Subsequently, the results of the two followup studies are debated. As two industries, namely the tobacco, and the integrated telecommunications services industry, were each represented by only one firm in the sample, no conclusions can be drawn from the results on these industries. Interpretations of these results are, therefore, not presented, as they would not be meaningful.

6.1. Main study

6.1.1. Descriptive statistics

The descriptive results of this study show that, contrary to what is reported in the literature, there was no continuous increase in the quantity of reporting and no continuous increase in the number of standalone CSR reports over time. However, this study only looked at the developments over

[13]The p-value indicated by the Hausman test (Prob > chi2 = 0.2971) was above the critical value of 0.05.

[14]The p-value indicated by the Hausman test (Prob > chi2 = 0.0839) was above the critical value of 0.05.

[15]An overview of these results can be found in Appendix A4. The models were, again, chosen based on the results of the Hausman tests.

Table 20: Level combinations on MBR: Random-effects regression results

	Coefficient	z	P>|z|
CombGroup			
2	-.3917638	-1.82	0.068
3	-.1982372	-0.67	0.506
4	-.0598091	-0.21	0.832
Industry			
10	6.075006	4.08	0.000
11	5.94829	2.29	0.022

Dependent variable: MBR
R^2=.5189
Number of obs.=392

Table 21: Level combinations on MBR: Mixed-effects regression results

	Coefficient	z	P>|z|
CombGroup			
2	-.4167528	-1.96	0.050
3	-.2169415	-0.74	0.462
4	-.0736784	-0.26	0.792
Industry			
9	3.625125	2.25	0.025
10	6.07751	5.03	0.000
11	6.166322	2.95	0.003

Dependent variable: MBR
Number of obs.=392

the past eight years. Stakeholders have been exerting pressure on companies to report on their CSR commitment for much longer than this. It is, therefore, possible that most companies had already reacted to this pressure and adapted their activities accordingly before 2011. This would explain the lack of an increase.

As mentioned in chapter 2, the MDGs, and therefore six of the examined CSR topics, were already established before 2011. The SDGs were first explicitly mentioned in the reports on the year 2016, which were published in 2017. There was no strong increase in hit counts for the newly introduced CSR topics in that year. The fact that these topics were already identified in the reports of the preceding years indicates that the keywords used to create the index for CSR communication breadth were chosen broadly enough to identify the content of the CSR topics, regardless of whether they were explicitly mentioned in connection with the SDGs. This assures the usefulness of this index to a certain degree.

The slight but continuous increase in CSR communication breadth suggests that firms have recognized the need to address a higher number of CSR topics in order to satisfy the demands of a wider range of stakeholders. In 2016, the first reports after the Volkswagen scandal were published. It is possible that companies were somewhat intimidated at the time and did not dare to make their CSR communication too broad. This would explain the sudden decrease in the breadth of communication in that year.

While the communication of many topics increased steadily over the observation period, the communication of some topics had its ups and downs in certain years. This supports the assumption that the content and focus of CSR are continuously changing, most likely in line with the altering interests and priorities of stakeholders and shareholders. these priorities are often influenced by trends or specific events (e.g., the increased media coverage of climate change, the scandal surrounding Harvey Weinstein, or the most recent “Fridays for Future” marches). (Schaltegger et al., 2006) A general decline in the communication of most topics could be observed in 2016. Few topics defied this decline. These topics may be considered as most important, defying trends in communication strength. Topics whose communication experienced a particularly sharp decline during such a period may be considered least important. Therefore, they are the first to fall victim to the trend of communication cutbacks.

The energy and pharma industry reported most broadly on their CSR activity, while the real estate industry reported the least broadly. The nature of the energy industry is already tied to a CSR topic, which they will naturally communicate extensively on. The energy industry, depending on the focus of their business, can either be perceived as conductive or constructive to sustainability. Therefore, firms might want

to elaborate on their responsible business practices, clarifying their commitment to CSR. The pharma industry suffers from a controversial reputation, often being described as "cutthroat". The products of the firms in this industry are often ingested directly and are crucial to the health of consumers. Therefore, being seen as reliable and responsible, is very important to this industry, which might explain their broad CSR reporting.

The industries which reported the most with regard to reporting quantity were the oil and gas, and the insurance industry. The oil and gas industry suffers from a bad connotation, being prone to scandals like oil spills and fracking controversies. Therefore, it makes sense that firms from such an industry would try to improve their image with stakeholders and lower their riskiness perceived by shareholders through long and detailed CSR reporting. The insurance industry would very much profit from a reliable and responsible image, as people, to a certain degree, trust this industry with their physical and financial livelihood. Literal contracts are being made between firms in this industry and its customers. As explained in the previous, a moral contracting partner is a desirable contracting partner. Companies in this industry may, therefore, make the extra effort of long CSR reports to establish the reputation of holding themselves to high moral standards. The industries with the least voluminous CSR reporting were the food processing and retailing, the pharma, and the real estate industry. The food production and retailing industry might consider their products as too essential to feel the need to excessively report on their CSR activities. As the pharma industry scores very high on the breadth of communication, it could be argued that the firms in this industry prioritize the content of reporting over the quantity of reporting or simply feel that it is better to communicate their CSR in a broad but concise manner. Again, the real estate industry is among the industries that communicate the least. It, therefore, appears that CSR reporting is generally considered as not too important in this industry.

The topics that were communicated the most were “peace, justice, and strong institutions”, “affordable and green energy”, “climate action”, and “responsible consumption and production”. Climate change has been a very prominent topic over the past years. Both “affordable and green energy” and “climate action” are closely tied to this topic, while “responsible consumption and production” is also linked to it to some degree. Even if the topic of “peace, justice, and strong institutions” seems to be on a superordinate level to that of companies, issues such as human rights, corruption, and bribery were increasingly addressed in company reports. These are problems that directly affect companies or even emanate from them.

The topics that were communicated the least were “life below water”, “reduced inequalities”, “no poverty”, and “life on land”. The fact that poverty alleviation has been placed as the first SDG was interpreted by some researchers (e.g., Hoque et al., 2018) as an indication that this topic was of the most importance. Companies do not seem to agree with this interpretation. They may see such a topic as not their responsibility but that of policymakers. The same might be the case for the topic of "reduced inequalities", which deals with the reduction of inequalities within and among countries. Companies seem to give high priority to issues related to climate change as they seem to communicate them extensively. The topic of "life on land", which deals with the protection of forest areas and land degradation, may, therefore, seem redundant or too specific to them, as they have already addressed climate change by discussing other topics. None of the industries in our sample had a direct connection to the ocean. Therefore, the topic of "life below water" may not have been particularly relevant for the companies investigated. If companies in, for example, the shipping industry had been investigated, the result might have been different.

Companies with a higher number of employees communicated in higher quantity and more broadly than smaller firms. With greater firm size, usually comes a greater amount of both internal and external transactions (Godfrey et al., 2009; Kimberly, 1976). Due to the more considerable breadth and volume of a company’s transactions, it will probably have an impact on more CSR topics, which can then be reported. Besides, CSR reporting can be a significant cost factor (Oberndorfer, Schmidt, Wagner, & Ziegler, 2013). It could be argued that large companies simply have more resources to spend on CSR reporting.

The industries with the highest ESG scores were the real estate, the energy, and the insurance industry. While the energy industry also communicates extensively, the real estate industry was among those industries that communicated the least (in breadth and quantity). It makes perfect sense for industries like the insurance, and energy industry, that scores high in the ESG assessment to also generously report its efforts. However, it is surprising to see that an industry that has been found to act responsibly would not want to report on its CSR commitments.

Personal products are products with high visibility that are directly used by the consumer. This puts them under constant surveillance. Consequently, the motivation to manage their CSR commitments could be accordingly high. This could explain the high frequency in which firms in the personal products industry won sustainability awards.

6.1.2. Regression results

The results did not differ significantly between the fixed-effects, random-effects, and mixed-effects models. In a few cases, the p-values differed slightly between the models, while the direction of the effects remained the same. This is due to the differences in the way the models transform the data (Baltagi, 2005). In these cases, the influences were interpreted as tendencies. Overall, the results can be considered to be satisfactorily robust, and can, therefore, be interpreted and discussed.

The results of this thesis suggest that CSR commitment has no significant influence on accounting-based CFP. This could be explained by the fact that investments in CSR commitment are often not immediately visible, and most likely do not affect accounting-based CFP instantaneously. The effect

of those investments might, therefore, only become apparent after several years (K.-H. Kim et al., 2018). Interestingly, while not being significant, the effect of CSR commitment on accounting-based CFP was actually negative. The attempt to satisfy all the interests of different stakeholder groups can lead to an inefficient use of resources and a decrease in CFP in the long run (Aupperle, Carroll, & Hatfield, 1985). Although irresponsible corporate behavior can have a negative influence on the relationship of firms and their stakeholders and might diminish a firm's public reputation, it can still generate value for the firm as it is often associated with cost-saving (Mani & Wheeler, 1998; Strike, Gao, & Bansal, 2006; Tang, Qian, Chen, & Shen, 2015). An example of this would be the use of nonrefined oil, causing environmental pollution, while saving operating costs at the same time (M. Kim & Kim, 2019). Therefore, the costs of CSR commitment could put socially responsible companies at an economic disadvantage (Ullmann, 1985), explaining the results.

The results also revealed that CSR commitment has a negative influence on market-based CFP[16]. Considering the small sample size, the p-value of the influence of winning a sustainability award on market-based CFP was not too far from being statistically significant. There-fore, the effect deserves to be mentioned. Previous research shows that firms have been punished on the stock market for the inclusion in a sustainability stock index. Investors might suspect that these firms spent too much on CSR, and that, as a consequence, more profitable projects might be overlooked or underfunded by the firm. (Oberndorfer et al., 2013) Being awarded a sustainability prize may give the same impression to stakeholders. They might fear that a firm is not distributing its resources economically, making the firm a less attractive investment option.

CSR communication has an influence on both accounting-based and market-based CFP. The results show a clear tendency that the publication of a standalone CSR report is perceived positively by stakeholders, while shareholders seem to be indifferent towards it. CSR reporting is not a one-time decision but requires continuous assessment and management of a firm's practices with regard to CSR (Székely & Knirsch, 2005). Stakeholders might, therefore, see a standalone CSR report as a clear sign that a firm actively thinks about and evaluates its CSR commitments. This could be different for shareholders, as they draw essential information from the annual reports of companies, which are later incorporated into their investment decisions. It is said that shareholders are the main audience for annual financial reports (Murray et al., 2006). If information about CSR reports is now incorporated into what were previously purely financial reports, this information would automatically be read by shareholders. Therefore, it probably makes little difference to shareholders whether a standalone CSR report is published or not.

High levels of CSR communication quantity affect both accounting- and market-based CFP negatively. Excessive quantities of CSR communication might not always be perceived as a sign of transparency by stakeholders and shareholders, but rather raise suspicion about the sincerity of the reporting and the factual correctness of the reported data (Dando & Swift, 2003; Doane & Potts, 2000). Overly lengthy CSR reports could, therefore, cause the opposite of the desired effect, making the firm look insincere and untrustworthy. This adversely affects a firm's relationship with its stakeholders, which might lead to stakeholders withdrawing their support and resources from the firm. Untrustworthy firms are also perceived to be a riskier investment option (Alexander & Buchholz, 1978). Financial market participants will, therefore, be more reluctant to invest in such a company. All of these reactions would lead to weaker CFP.

The breadth of CSR communication has no influence on accounting-based CFP. Stakeholders might already be content with the mere existence of a CSR reporting and might not deal with the contents of the reporting in great depth. The breadth of the CSR topics that are being addressed in the reports would, therefore, not be recognized, explaining the non-existent effect.

In contrast to accounting-based CFP, addressing a broad range of CSR topics has a positive influence on a firm's market-based CFP. Financial market participants are on a constant quest for information that can be incorporated in their risk assessment of risks and decision-making processes (Malsch, 2013). CSR is considered to play a part of a firm's risk management (Michelon et al., 2020). Firms with a broad CSR communication appear to act responsibly in many different areas. This minimized the risk of misconduct in one CSR area, which would, if discovered, result in a public scandal. Also, broad CSR communication gives the impression that the firm in question is already adhering to CSR standards, which are above the legal requirements in many different business areas. This minimizes the risk of the firm being hurt by any new legislation or regulations concerning CSR standards. Scandals or the burden of adjusting operations or production to new laws would harm a firm financially, putting shareholders at a disadvantage. Companies that present themselves as exemplary with regard to CSR would, therefore, be considered a more attractive investment option.

It is often assumed that shareholders are not only the main audience of financial reports, but also of CSR reports, and that other stakeholders seldomly pay attention to these disclosures (Murray et al., 2006). The results of this thesis suggest that CSR reports are also perceived and evaluated by stakeholders. However, it is interesting to see that the presence of a standalone CSR report has an impact on accounting-based CFP, while it is the content of the reporting that influences market-based CFP. This reaffirms the assumption made before that that shareholders deal with the reports in more depth, while stakeholders react to more noticeable and quickly assessable parameters.

[16] Although ESGScore technically had a positive influence on MBR, it was so small and statistically insignificant that it will not be considered.

6.2. Follow-up studies

6.2.1. Follow-up study 1

As CSR communication breadth did not have an influence on accounting-based CFP, it was no surprise that no single CSR topic had a significant influence on accounting-based CFP either. This supports the assumption that the content of CSR reports is not particularly relevant to stakeholders or that they simply do not read the reports in detail. Many stakeholder groups, like customers or NGOs, have no direct financial stake in the companies. If a company evokes a scandal, in most cases, it is easy for customers to just switch to an alternative brand. The consequences of such a scandal would be much more severe for shareholders. Therefore, stakeholders might just not inform themselves about the activities of companies in the same vigorous manner as shareholders.

Three CSR topics had a positive influence on market-based CFP. These topics were “reduced inequalities in and among countries”, “gender equality”, and “affordable and clean energy”. The communication of the topic of “reduced inequalities in and among countries” had the most considerable influence. While not many firms communicated on this topic, the extra effort to take up this ambitious goal seemed to resonate well with shareholders.

As gender equality has been a critical topic in recent years, it was not surprising that shareholders would act favorably towards firms that show concern and engagement in this issue. Women’s marches and the “#metoo” movement are only a few prominent examples of recent initiatives to improve women’s standing in society and the business world. Showing support of this movement is almost expected from firms nowadays and rewarded by shareholders. Scandals revolving around the mistreatment of women are no longer affordable for companies in this age. As was seen in the case of "The Weinstein Company", multimillion-dollar compnies can be plunged into insolvency by serious and confirmed allegations. Not to take a clear position on or to lack commitment in this problem could lead to a reputational loss of a company, while being vocal on this issue seems to improve reputation.

The topic of the energy system transformation has been very relevant for some time now, especially due to the decision of some European countries to participate in the nuclear power phaseout in recent years. Therefore, it is not surprising that shareholders seem to value a firm’s engagement in this area. Renewable energy sources are often considered as a way of saving costs and consequently improving CFP in the long term. It is, therefore, also in the interest of shareholders when a company addresses this issue.

The topic of "industry, innovation, and infrastructure" had a negative influence on market-based CFP. The goal of this topic is to “build resilient infrastructure, promote inclusive and sustainable industrialization, and foster innovation” (United Nations, 2016, p. 17). This ambitious topic could be seen as too high a cost factor, explaining the negative impact on market-based CFP. Also, it usually takes some time for innovations to break even financially. The initial reactions to such investments might, therefore, be a bit skeptical.

6.2.2. Follow-up study 2

The descriptive results of this study show that there is indeed a large number of firms that display a communication-behavior gap when it comes to their CSR activities. Rules and regulations on CSR reporting, as well as the verification of its content, are evidently not strict enough as firms do not seem to be scared of being discovered and punished for their behavior.

No constellation of the displayed levels of commitment and communication had a statistically significant influence on accounting-based CFP. Displaying low levels of both CSR commitment and CSR communication was the only combination that showed a positive, yet still insignificant, influence. Looking at the levels separately, both displaying a high level of CSR commitment or a high level of CSR communication had a negative but statistically insignificant influence on accounting-based CFP. Although they are not significant, the results indicate that CSR commitment and CSR communication both represent high costs for a company that may not be offset by the benefits of good stakeholder relationships. The results also suggest that stakeholders do not seem to be aware of any discrepancies between CRS commitment and CSR communication of firms. This, again, supports the assumption that stakeholders do not concern themselves with the CSR actions of firms in much depth.

Displaying a high level of CSR communication while displaying only low levels of CSR commitment has a negative influence on market-based CFP. This indicates that shareholders see behind the facade of extensive CSR communication and punish those firms that, intentionally or unintentionally, try to deceive the public. This can be attributed to the fact that shareholders are more deeply concerned with the CSR activities of companies, as they are financially involved with them, and any misconduct on the side of the companies can come at a high cost for shareholders.

7. Limitations and further research opportunities

Although the results of this study are intriguing, as with any research, some limitations must be evaluated. These limitations raise the possibility for further research, which will also be elaborated.

There are some limitations concerning the choice and construction of the variables. Compromises had to be made on one independent and one control variable due to access and availability difficulties. The number of articles on a firm on the Handelsblatt website does not measure the same thing as marketing expenses, with regard to firm visibility. Firms spend money on marketing to make themselves appear in a favorable light, while their portrayal in articles is mostly out of the firms’ hands. A high visibility in the economic media does not necessarily mean a good thing for the firm.

There are also some concerns about the ESG Combined Score. While the Thomson Reuters claims to use the same method as ISS ESG to investigate firms and compose the final score, when looking at the few ISS ESG scores which

were made public, considerable differences could be found between the final scores. Another concern with the ESG Combined Score is that, while the methodology behind this score is thorough and its coverage wide, it might not be completely appropriate as a metric. ESG information is often published in the non-financial part of reporting. As firms' ESG disclosures can vary considerably, ESG data might not be generally comparable (Dinçer & Yüksel, 2018). While the rating's methodology goes beyond the review of voluntary disclosure, Thomson Reuters does use some information that has been published by the firms under question themselves (Thomson Reuters, 2013), and it is not clear how strongly the rating is based on this disclosure (Beck et al., 2018).

It is also not clear whether ESG scores, like the Thomson Reuters ESG Combined Score, are actually comprehensible to stakeholders and shareholders (Luo et al., 2015), and, therefore, how useful they are as a variable. It should, therefore, be considered whether ESG scores, which are so often used in research as a proxy for CSR commitment, represent effective indicators for CSR commitment. Also, as different operationalization of CSR commitment measures has shown to lead to different results about the relationship between CSR and CFP (Busch & Hoffmann, 2011), it should be examined if and how the results of this study change if the variable representing CSR commitment is based on alternative sources for CSR validation, which do not rely on an ESG database (Singal, 2014).

There were also some limitations concerning the statistical method. While the values for R^2 of the random-effects regression models were decent, the R^2 values of the fixed-effects models were very small. Therefore, the variables in the model do not explain a big proportion of the variation in the dependent variable. This could indicate that important independent variables have were omitted. There has long been a critical debate about the information content and usefulness of R^2. King (1986) argues that R^2 is a poor measure and highly misused. Therefore, this limitation might not be too severe. However, future research should pick up on this and include new variables that the researchers consider significant. Such an inclusion of further variables, like, for example, R&D investment, might also lead to the relationship between CSR and CFP being no longer statistically significant (McWilliams & Siegel, 2000).

Additionally, the results of the random-effects, fixed-effects models, and mixed-effects models displayed a few discrepancies, mainly concerning the p-values. Therefore, the influence of some variables had to be interpreted as tendencies. Future research could pick up on this and see whether the p-values are still different when the sample size is increased.

There might also be a causality issue in this study. CSR commitment and communication might not just be a predictor but also a consequence of CFP, forming a sort of circle (Surroca, Tribó, & Waddock, 2010; Waddock & Graves, 1997). In future research, the models presented in this paper could be modified accordingly and then tested again.

Furthermore, the eight-year period of this study was not long enough to produce generalizable results. Therefore, one should be cautious when deriving inference from these findings (Choi et al., 2010). The same limitation arises from the relatively small sample size. A followup study over a longer time frame with a bigger sample size should be conducted to see if the results can be replicated. Also, all the firms in the sample belonged to the 80th percentile of their industry with regard to market capitalization and were publicly traded. The results might, therefore, not apply to small and privately-owned companies. It would be interesting to see whether the results are different for firms with a smaller market capitalization.

Finally, the present study could be replicated on a much larger scale to shed further light on how different stakeholder groups perceive CSR. Different stakeholders display differing expectations concerning a firm's behavior (Fombrun et al., 2000). Some stakeholder groups might be more prominent in one industry than in the other. Therefore, the effect of different types of CSR activities on corporate reputation and, consequently, CFP, might differ between industries (Rodriguez-Fernandez, 2016). It would be interesting to repeat this study with different samples, each containing companies from one industry. This would make it possible to assess whether the stakeholders of companies in one industry have different CSR priorities than those of another industry.

8. Implications

8.1. Theoretical contribution

The relationship between CSR and CFP has been the center of debate in research for decades now. Many studies draw on stakeholder theory to explain this relationship (Donaldson, 1999; Freeman, 2010; T. M. Jones, 1995). According to instrumental stakeholder theory, firms with outstanding social performance also have the tendency to perform better financially (Baron, 2001; T. M. Jones, 1995). This study adds to the literature on instrumental stakeholder theory, showing that companies that respond to the desire of their stakeholders and shareholders for more CSR-related information can benefit from an increased CFP. Many studies in this research area have failed to distinguish between CSR commitment and CSR communication when quantifying CSR and treated these activities synonymously. The results of this thesis make it clear that there are often great discrepancies between the levels of CSR commitment and CSR communications. These findings support the literature on the communication-behavior gap, although they do not support the notion that this gap has widened over the past couple of years. The results also partly support the literature on the risks of allegations of greenwashing, and the damage they do on corporate reputation. They do, however, expand on this by revealing that untrustworthiness is primarily based on the quantity of communication, while the breadth of communication seems to be beneficial.

While the role of the quantity of CSR reporting has been researched widely in the existing investment literature (Dhaliwal et al., 2012), the content of this reporting

has not received an equal amount of attention (Harjoto & Jo, 2015). This work contributes to the neglected re-search area. The index for the breadth of CSR communication, which was created especially for this work, also offers a tool for the assessment of the content of CSR reports in future re-search.

8.2. Practical implications

The results of this thesis provide interesting and important practical implications. Many prior studies have debated the question of whether firms should invest in CSR commitments and the reporting of these (Orlitzky et al., 2003). While no clear statements can be made about the benefits of CSR commitment, conclusions can be drawn about CSR communication.

The preparation of CSR reports is expensive and time consuming, and managers are often not entirely sure if these publications evoke the aspired information dissemination (García-Sánchez et al., 2019; O'dwyer, 2002). The results of this work indicate that CSR reports are indeed perceived and valued by stakeholders and shareholders. The literature often cites shareholders as the most important audience for annual and CSR reports (e.g., Murray et al., 2006). Although the results suggest that shareholders are more interested in the content of these reports, it is evident that these reports also influence the other stakeholders. Companies can therefore be advised to continue preparing and publishing CSR reports.

The high percentage of companies that communicate extensively about CSR activities but display little actual CSR commitment shows the lack of verification of CSR reporting. Urgent work should therefore be done on regulations that control what and how much companies publish with regard to their CSR commitment. While CSR communication triggers a reaction in stakeholders and shareholders, they do not seem to be aware of CSR commitment. CSR commitment comes at a high cost for the company, but its manifestations are often barely visible and difficult to assess. New rules and regulations regarding CSR communication must guarantee that only companies with a high level of commitment are able to engage in high level communication, so that stakeholders and shareholders can rely on the information presented to them in their assessment of firms. However, it is reassuring that at least shareholders seem to see through excessive communication on CSR issues on the part of companies. Companies are therefore urged to keep to the truth when communicating CSR issues. Nevertheless, these results should not merely scare dishonest companies into refraining from swindling about their sense of responsibility. They should also serve as an instruction for honest companies on how to communicate their CSR commitment effectively, as this seems to pose a challenge to many (Adams, Potter, Singh, & York, 2016; Qiu, Shaukat, & Tharyan, 2016). Producing shorter CSR reports with a higher information content could not only improve the image of the company among stakeholders and shareholders, but also possibly save the company time and money. However, it is questionable whether, based on the results of this work, firms should be advised to communicate more on or even exclusively specific CSR. Companies should rather get an overview of which CSR topics are particularly important to their stakeholders and shareholders. Based on this information, efforts and resources can then be allocated to the individual topics.

9. Conclusion

The main objective of the presented work was to examine whether CSR commitment and CSR communication, separately, have an influence on CFP and if their influences differ in direction and strength. 656 annual and CSR reports were analyzed to create precise variables to test these relationships. Through this accurate and extensive data collection, this thesis contributes significantly to the debate on the relationship between CSR and CFP. A panel data analysis on 49 European firms over an observation period of eight years was then conducted. The results of this analysis revealed that while CSR commitment has no impact on CFP, the relationship between CSR communication and CFP is somewhat more complicated. While standalone CSR reports and broad communication on CSR topics are positively received by stakeholders and shareholders, respectively, both groups are opposed to high quantities of CSR reporting.

The results of the main and the followup studies suggest that shareholders concern themselves with the CSR actions of companies more deeply than stakeholders. The communication of different CSR topics also influences market-based CFP differently. In addition, displaying high levels of communication while displaying low levels of commitment has a negative influence on market-based CFP. This indicates that, while shareholders appreciate CSR reporting, they do not appreciate being deceived. Overall, the results of the main, as well as the followup studies, suggest CSR reporting can help a company to satisfy its stakeholders and shareholders and, therefore, improve CFP, but firms have to dread carefully not to be perceived as dishonest.

The CSR activities of companies are increasingly moving into the focus of stakeholder groups. While these activities are known to involve costs for companies, they are also associated with higher CFP as they are a means of improving stakeholder relationships. The clear categorization of CSR activities into commitment and communication and the separate examination of accounting-based and market-based CFP in this work revealed new insights into this relationship between CSR and CFP. Relying on the findings of this thesis, the two initial questions have to be answered: Does it pay to actually be good? Not really. But does it pay to deceive? Definitely not!

References

Abbott, W. F., & Monsen, R. J. (1979). On the measurement of corporate social responsibility: Self-reported disclosures as a method of measuring corporate social involvement. *Academy of Management Journal*, *22*(3), 501–515.

Ackoff, R. L. (1974). *Redesigning the Future: A System Approach to Societal Problems*. John Wiley & Sons.

Adams, C. A. (2002). Internal organisational factors influencing corporate social and ethical reporting: Beyond current theorising. *Accounting, Auditing & Accountability Journal*.

Adams, C. A., Potter, B., Singh, P. J., & York, J. (2016). Exploring the implications of integrated reporting for social investment (disclosures). *The British Accounting Review*, *48*(3), 283–296.

Alexander, G. J., & Buchholz, R. A. (1978). Corporate social responsibility and stock market performance. *Academy of Management Journal*, *21*(3), 479–486.

Al-Tuwaijri, S. A., Christensen, T. E., & Hughes Ii, K. (2004). The relations among environmental disclosure, environmental performance, and economic performance: a simultaneous equations approach. *Accounting, Organizations and Society*, *29*(5-6), 447–471.

Ameer, R., & Othman, R. (2012). Sustainability practices and corporate financial performance: A study based on the top global corporations. *Journal of Business Ethics*, *108*(1), 61–79.

Anderson, J. C., & Frankle, A. W. (1980). Voluntary social reporting: An iso-beta portfolio analysis. *Accounting Review*, 467–479.

Argandoña, A., & von Weltzien Hoivik, H. (2009). Corporate social responsibility: One size does not fit all. Collecting evidence from Europe. *Journal of Business Ethics*, *89*(3), 221–234.

Aupperle, K. E., Carroll, A. B., & Hatfield, J. D. (1985). An empirical examination of the relationship between corporate social responsibility and profitability. *Academy of Management Journal*, *28*(2), 446–463.

Auret, C., & Sinclaire, R. (2006). Book-to-market ratio and returns on the JSE. *Investment Analysts Journal*, *35*(63), 31–38.

Baboukardos, D. (2018). The valuation relevance of environmental performance revisited: The moderating role of environmental provisions. *The British Accounting Review*, *50*(1), 32–47.

Baltagi, B. H. (2005). *Econometric analysis of panel data*. 3rd Edition England JW & Sons.

Baltagi, B. H., Bresson, G., & Pirotte, A. (2003). Fixed effects, random effects or Hausman–Taylor?: A pretest estimator. *Economics Letters*, *79*(3), 361–369.

Bansal, P., & Clelland, I. (2004). Talking trash: Legitimacy, impression management, and unsystematic risk in the context of the natural environment. *Academy of Management Journal*, *47*(1), 93–103.

Barkemeyer, R., Holt, D., Preuss, L., & Tsang, S. (2014). What happened to the 'development'in sustainable development? Business guidelines two decades after Brundtland. *Sustainable Development*, *22*(1), 15–32.

Barnett, M. L., & Salomon, R. M. (2006). Beyond dichotomy: The curvilinear relationship between social responsibility and financial performance. *Strategic Management Journal*, *27*(11), 1101–1122.

Baron, D. P. (2001). Private politics, corporate social responsibility, and integrated strategy. *Journal of Economics & Management Strategy*, *10*(1), 7–45.

Basu, K., & Palazzo, G. (2008). Corporate social responsibility: A process model of sensemaking. *Academy of Management Review*, *33*(1), 122–136.

Bebbington, J., Larrinaga, C., & Moneva, J. M. (2008). Corporate social reporting and reputation risk management. *Accounting, Auditing & Accountability Journal*.

Beck, C., Frost, G., & Jones, S. (2018). CSR disclosure and financial performance revisited: A cross-country analysis. *Australian Journal of Management*, *43*(4), 517–537.

Belal, A. R., & Cooper, S. (2011). The absence of corporate social responsibility reporting in Bangladesh. *Critical Perspectives on Accounting*, *22*(7), 654–667.

Belkaoui, A. (1976). The impact of the disclosure of the environmental effects of organizational behavior on the market. *Financial Management*, 26–31.

Bénabou, R., & Tirole, J. (2010). Individual and corporate social responsibility. *Economica*, *77*(305), 1–19.

Berliner, D., & Prakash, A. (2015). "Bluewashing" the Firm? Voluntary Regulations, Program Design, and Member Compliance with the United Nations Global Compact. *Policy Studies Journal*, *43*(1), 115–138.

Bernardi, C., & Stark, A. W. (2018). Environmental, social and governance disclosure, integrated reporting, and the accuracy of analyst forecasts. *The British Accounting Review*, *50*(1), 16–31.

Berthelot, S., Coulmont, M., & Serret, V. (2012). Do investors value sustainability reports? A Canadian study. *Corporate Social Responsibility and Environmental Management*, *19*(6), 355–363.

Bird, R., Hall, A. D., Momentè, F., & Reggiani, F. (2007). What corporate social responsibility activities are valued by the market? *Journal of Business Ethics*, *76*(2), 189–206.

Borgers, A., Derwall, J., Koedijk, K., & Ter Horst, J. (2015). Do social factors influence investment behavior and performance? Evidence from mutual fund holdings. *Journal of Banking & Finance*, *60*, 112–126.

Bowman, E. H., & Haire, M. (1975). A strategic posture toward corporate social responsibility. *California Management Review*, *18*(2), 49–58.

Brammer, S. J., & Pavelin, S. (2006). Corporate reputation and social performance: The importance of fit. *Journal of Management Studies*, *43*(3), 435–455.

Brigham, E. F., & Ehrhardt, M. C. (2013). *Financial Management: Theory & Practice (Book Only)*. Cengage Learning.

Brown, K. A., Willis, P. G., & Prussia, G. E. (2000). Predicting safe employee behavior in the steel industry: Development and test of a sociotechnical model. *Journal of Operations Management*, *18*(4), 445–465.

Brown, N., & Deegan, C. (1998). The public disclosure of environmental performance information—a dual test of media agenda setting theory and legitimacy theory. *Accounting and Business Research*, *29*(1), 21–41.

Brown, S. P. (1996). A meta-analysis and review of organizational research on job involvement. *Psychological Bulletin*, *120*(2), 235.

Brundtland, G. (1987). Our Common Future: Report of the 1987 World Commission on Environment and Development. *United Nations, Oslo*, *1*, 59.

Busch, T., & Hoffmann, V. H. (2011). How hot is your bottom line? Linking carbon and financial performance. *Business & Society*, *50*(2), 233–265.

Cahan, S. F., De Villiers, C., Jeter, D. C., Naiker, V., & Van Staden, C. J. (2016). Are CSR disclosures value relevant? Cross-country evidence. *European Accounting Review*, *25*(3), 579–611.

Carroll, A., Brown, J., & Buchholtz, A. (2017). *Business & society: Ethics, sustainability & stakeholder management (2nd)*. Mason: Cengage.

Castelló, I., Morsing, M., & Schultz, F. (2013). Communicative dynamics and the polyphony of corporate social responsibility in the network society. *Journal of Business Ethics*, *118*(4), 683–694.

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

Chiu, S.-C., & Sharfman, M. (2011). Legitimacy, visibility, and the antecedents of corporate social performance: An investigation of the instrumental perspective. *Journal of Management*, *37*(6), 1558–1585.

Cho, C. H., & Patten, D. M. (2007). The role of environmental disclosures as tools of legitimacy: A research note. *Accounting, Organizations and Society*, *32*(7-8), 639–647.

Cho, C. H., Roberts, R. W., & Patten, D. M. (2010). The language of US corporate environmental disclosure. *Accounting, Organizations and Society*, *35*(4), 431–443.

Choi, J.-S., Kwak, Y.-M., & Choe, C. (2010). Corporate social responsibility and corporate financial performance: Evidence from Korea. *Australian Journal of Management*, *35*(3), 291–311.

Claasen, C., & Roloff, J. (2012). The link between responsibility and legitimacy: The case of De Beers in Namibia. *Journal of Business Ethics*, *107*(3), 379–398.

Clarkson, M. B. E. (1995). A Stakeholder Framework for Analyzing and Evaluating Corporate Social Performance. *The Academy of Management Review*, *20*(1), 92–117. Retrieved from http://www.jstor.org/stable/258888

Clarkson, P. M., Fang, X., Li, Y., & Richardson, G. (2013). The relevance of environmental disclosures: Are such disclosures incrementally informative? *Journal of Accounting and Public Policy*, *32*(5), 410–431.

Clarkson, P. M., Li, Y., Richardson, G. D., & Vasvari, F. P. (2008). Revisiting the relation between environmental performance and environmen-

tal disclosure: An empirical analysis. *Accounting, Organizations and Society*, *33*(4-5), 303–327.
Coldwell, D. A., Billsberry, J., Van Meurs, N., & Marsh, P. J. (2008). The effects of person–organization ethical fit on employee attraction and retention: Towards a testable explanatory model. *Journal of Business Ethics*, *78*(4), 611–622.
Commission of the European Communities. (2002). *Corporate Social Responsibility: A business contribution to Sustainable Development*.
Coombs, W. T., & Holladay, S. J. (2011). *Managing corporate social responsibility: A communication approach*. John Wiley & Sons.
Cormier, D., & Magnan, M. (2014). The impact of social responsibility disclosure and governance on financial analysts' information environment. *Corporate Governance*.
Creyer, E. H. (1997). The influence of firm behavior on purchase intention: do consumers really care about business ethics? *Journal of Consumer Marketing*.
Dando, N., & Swift, T. (2003). Transparency and assurance minding the credibility gap. *Journal of Business Ethics*, *44*(2), 195–200.
Daub, C.-H. (2007). Assessing the quality of sustainability reporting: an alternative methodological approach. *Journal of Cleaner Production*, *15*(1), 75–85.
Davenport, K. (2000). Corporate citizenship: A stakeholder approach for defining corporate social performance and identifying measures for assessing it. *Business & Society*, *39*(2), 210–219.
Deegan, C., Rankin, M., & Tobin, J. (2002). An examination of the corporate social and environmental disclosures of BHP from 1983-1997: A test of legitimacy theory. *Accounting, Auditing & Accountability Journal*.
De Klerk, M., De Villiers, C., & Van Staden, C. (2015). The influence of corporate social responsibility disclosure on share prices: Evidence from the United Kingdom. *Pacific Accounting Review*.
Delmas, M. A., & Burbano, V. C. (2011). The drivers of greenwashing. *California Management Review*, *54*(1), 64–87.
Dess, G. G., Ireland, R. D., & Hitt, M. A. (1990). Industry effects and strategic management research. *Journal of Management*, *16*(1), 7–27.
Dhaliwal, D. S., Radhakrishnan, S., Tsang, A., & Yang, Y. G. (2012). Nonfinancial disclosure and analyst forecast accuracy: International evidence on corporate social responsibility disclosure. *The Accounting Review*, *87*(3), 723–759.
Dinçer, H., & Yüksel, S. (2018). *Handbook of research on managerial thinking in global business economics*. IGI Global.
Doane, D., & Potts, R. (2000). *Corporate spin: The troubled teenage years of social reporting*. New Economics Foundations.
Donaldson, T. (1999). Making stakeholder theory whole. *Academy of Management Review*, *24*(2), 237–241.
Donaldson, T., & Dunfee, T. (1999). Ties that Bind. *Harvard Business School Press, Boston*.
Donaldson, T., & Preston, L. E. (1995). The stakeholder theory of the corporation: Concepts, evidence, and implications. *Academy of Management Review*, *20*(1), 65–91.
Doni, F., & Johannsdottir, L. (2020). Environmental Social and Governance (ESG) Ratings. *Climate Action*, 435–449.
Dunne, T., Roberts, M. J., & Samuelson, L. (1989). The growth and failure of US manufacturing plants. *The Quarterly Journal of Economics*, *104*(4), 671–698.
Elias, R. Z. (2004). The impact of corporate ethical values on perceptions of earnings management. *Managerial Auditing Journal*.
Elkington, J. (2004). Biting the bullet points. In *European Business Forum, Special Report—EBF on Corporate Social Responsibility*.
Erickson, T., & Whited, T. M. (2006). On the accuracy of different measures of q. *Financial Management*, *35*(3), 5–33.
Firer, S., & Williams, S. M. (2003). Intellectual capital and traditional measures of corporate performance. *Journal of Intellectual Capital*.
Flammer, C. (2015). Does corporate social responsibility lead to superior financial performance? A regression discontinuity approach. *Management Science*, *61*(11), 2549–2568.
Fombrun, C. J., Gardberg, N. A., & Barnett, M. L. (2000). Opportunity platforms and safety nets: Corporate citizenship and reputational risk. *Business and Society Review*, *105*(1).
Fonseca, A. (2010). How credible are mining corporations' sustainability reports? A critical analysis of external assurance under the requirements of the international council on mining and metals. *Corporate Social Responsibility and Environmental Management*, *17*(6), 355–370.
Frank, R. H. (1988). *Passions within reason: The strategic role of the emotions*. WW Norton & Co.
Freeman, R. E. (1984). *Strategic management: A stakeholder approach*. Pitman.
Freeman, R. E. (2010). Managing for stakeholders: Trade-offs or value creation. *Journal of Business Ethics*, *96*(1), 7–9.
Freudenreich, B., Lüdeke-Freund, F., & Schaltegger, S. (2019). A stakeholder theory perspective on business models: Value creation for sustainability. *Journal of Business Ethics*, *166*(1), 3–18.
Gamerschlag, R., Möller, K., & Verbeeten, F. (2011). Determinants of voluntary CSR disclosure: empirical evidence from Germany. *Review of Managerial Science*, *5*(2-3), 233–262.
García-Sánchez, I.-M., Hussain, N., Martínez-Ferrero, J., & Ruiz-Barbadillo, E. (2019). Impact of disclosure and assurance quality of corporate sustainability reports on access to finance. *Corporate Social Responsibility and Environmental Management*, *26*(4), 832–848.
Godfrey, P. C. (2005). The relationship between corporate philanthropy and shareholder wealth: A risk management perspective. *Academy of Management Review*, *30*(4), 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 Management Journal*, *30*(4), 425–445.
Greene, W. H. (2002). *Econometric Analysis (5th ed.)*. Prentice Hall.
GRI. (2011). Sustainability reporting guidelines version 3.1. *Amsterdam: GRI*.
Griffin, J. J., & Mahon, J. F. (1997). The corporate social performance and corporate financial performance debate: Twenty-five years of incomparable research. *Business & Society*, *36*(1), 5–31.
Guidry, R. P., & Patten, D. M. (2010). Market reactions to the first-time issuance of corporate sustainability reports: Evidence that quality matters. *Sustainability Accounting, Management and Policy Journal*.
Gutsche, G., & Ziegler, A. (2019). Which private investors are willing to pay for sustainable investments? Empirical evidence from stated choice experiments. *Journal of Banking & Finance*, *102*, 193–214.
Hahn, R., & Kühnen, M. (2013). Determinants of sustainability reporting: a review of results, trends, theory, and opportunities in an expanding field of research. *Journal of Cleaner Production*, *59*, 5–21.
Haigh, M., & Hazelton, J. (2004). Financial markets: a tool for social responsibility? *Journal of Business Ethics*, *52*(1), 59–71.
Hall, R. (1993). A framework linking intangible resources and capabiliites to sustainable competitive advantage. *Strategic Management Journal*, *14*(8), 607–618.
Hammond, K., & Miles, S. (2004). Assessing quality assessment of corporate social reporting: UK perspectives. In *Accounting Forum* (Vol. 28, pp. 61–79).
Haque, F. (2017). The effects of board characteristics and sustainable compensation policy on carbon performance of UK firms. *The British Accounting Review*, *49*(3), 347–364.
Harjoto, M. A., & Jo, H. (2015). 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.
Hooghiemstra, R. (2000). Corporate communication and impression management–new perspectives why companies engage in corporate social reporting. *Journal of Business Ethics*, *27*(1), 55–68.
Hooks, J., Coy, D., & Davey, H. (2002). The information gap in annual reports. *Accounting, Auditing & Accountability Journal*.
Hopenhayn, H. A. (1992). Entry, exit, and firm dynamics in long run equilibrium. *Econometrica: Journal of the Econometric Society*, 1127–1150.
Hoque, N., Rahman, A. R. A., Molla, R. I., Noman, A. H. M., & Bhuiyan, M. Z. H. (2018). Is corporate social responsibility pursuing pristine business goals for sustainable development? *Corporate Social Responsibility and Environmental Management*, *25*(6), 1130–1142.
Hörisch, J., Freeman, R. E., & Schaltegger, S. (2014). Applying stakeholder theory in sustainability management: Links, similarities, dissimilarities, and a conceptual framework. *Organization & Environment*, *27*(4), 328–346.
Illia, L., Zyglidopoulos, S. C., Romenti, S., Rodríguez-Cánovas, B., & del Valle Brena, A. G. (2013). Communicating corporate social respon-

sibility to a cynical public. *MIT Sloan Management Review*, *54*(3), 2.

Ioannou, I., & Serafeim, G. (2015). The impact of corporate social responsibility on investment recommendations: Analysts' perceptions and shifting institutional logics. *Strategic Management Journal*, *36*(7), 1053–1081.

IUCN, UNEP, & WWF. (1980). *World conservation strategy: Living resource conservation for sustainable development.*

Jackob, N. (2008). Credibility effects. *In W. Donsbach (Ed.), International Encyclopedia of Communication*, 1044–1047.

Jacobs, B. W., Singhal, V. R., & Subramanian, R. (2010). An empirical investigation of environmental performance and the market value of the firm. *Journal of Operations Management*, *28*(5), 430–441.

Jain, P., Vyas, V., & Chalasani, D. P. S. (2016). Corporate social responsibility and financial performance in SMEs: A structural equation modelling approach. *Global Business Review*, *17*(3), 630–653.

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.

Jones, E. E. (1964). *Ingratiation: A social psychology analysis*. Appleton-Century-Crofts.

Jones, T. M. (1995). Instrumental Stakeholder Theory: A Synthesis of Ethics and Economics. *The Academy of Management Review*, *20*.

Joyner, B. E., & Payne, D. (2002). Evolution and implementation: A study of values, business ethics and corporate social responsibility. *Journal of Business Ethics*, *41*(4), 297–311.

Kapoor, S., & Sandhu, H. (2010). Does it pay to be socially responsible? An empirical examination of impact of corporate social responsibility on financial performance. *Global Business Review*, *11*(2), 185–208.

Karpoff, J. M., Lott, J. R., Jr, & Wehrly, E. W. (2005). The reputational penalties for environmental violations: Empirical evidence. *The Journal of Law and Economics*, *48*(2), 653–675.

Kempf, A., & Osthoff, P. (2007). The effect of socially responsible investing on portfolio performance. *European Financial Management*, *13*(5), 908–922.

Kim, K.-H., Kim, M., & Qian, C. (2018). Effects of corporate social responsibility on corporate financial performance: A competitive-action perspective. *Journal of Management*, *44*(3), 1097–1118.

Kim, M., & Kim, Y. (2019). CSR and shareholder value in the restaurant industry: The roles of CSR communication through annual reports. *Cornell Hospitality Quarterly*, *60*(1), 69–76.

Kimberly, J. R. (1976). Organizational size and the structuralist perspective: A review, critique, and proposal. *Administrative Science Quarterly*, 571–597.

King, G. (1986). How not to lie with statistics: Avoiding common mistakes in quantitative political science. *American Journal of Political Science*, 666–687.

Klassen, R. D., & McLaughlin, C. P. (1996). The impact of environmental management on firm performance. *Management Science*, *42*(8), 1199–1214.

Kolk, A., & Perego, P. (2010). Determinants of the adoption of sustainability assurance statements: An international investigation. *Business Strategy and the Environment*, *19*(3), 182–198.

KPMG. (2011). *KPMG International Survey of Corporate Responsibility Reporting 2011*. Retrieved from https://www.upj.de/fileadmin/user_upload/MAINdateien/Aktuelles/Nachrichten/kpmg_reportingsurvey_2011.pdf

Laplume, A. O., Sonpar, K., & Litz, R. A. (2008). Stakeholder theory: Reviewing a theory that moves us. *Journal of Management*, *34*(6), 1152–1189.

Laskar, N., Chakraborty, T. K., & Maji, S. G. (2017). Corporate sustainability performance and financial performance: Empirical evidence from Japan and India. *Management and Labour Studies*, *42*(2), 88–106.

Laufer, W. S. (2003). Social accountability and corporate greenwashing. *Journal of Business Ethics*, *43*(3), 253–261.

Liesen, A., Figge, F., Hoepner, A., & Patten, D. M. (2017). Climate change and asset prices: are corporate carbon disclosure and performance priced appropriately? *Journal of Business Finance & Accounting*, *44*(1-2), 35–62.

Lo, S.-F., & Sheu, H.-J. (2007). Is corporate sustainability a value-increasing strategy for business? *Corporate Governance: An International Review*, *15*(2), 345–358.

Lock, I., & Seele, P. (2017). Measuring credibility perceptions in CSR communication: A scale development to test readers' perceived credibility of CSR reports. *Management Communication Quarterly*, *31*(4), 584–613.

Luo, X., Wang, H., Raithel, S., & Zheng, Q. (2015). Corporate social performance, analyst stock recommendations, and firm future returns. *Strategic Management Journal*, *36*(1), 123–136.

Lyon, T. P., & Maxwell, J. W. (2011). Greenwash: Corporate environmental disclosure under threat of audit. *Journal of Economics & Management Strategy*, *20*(1), 3–41.

Maak, T. (2008). Undivided corporate responsibility: Towards a theory of corporate integrity. *Journal of Business Ethics*, *82*(2), 353–368.

Mahoney, L. S., Thorne, L., Cecil, L., & LaGore, W. (2013). A research note on standalone corporate social responsibility reports: Signaling or greenwashing? *Critical perspectives on Accounting*, *24*(4-5), 350–359.

Malsch, B. (2013). Politicizing the expertise of the accounting industry in the realm of corporate social responsibility. *Accounting, Organizations and Society*, *38*(2), 149–168.

Mani, M., & Wheeler, D. (1998). In search of pollution havens? Dirty industry in the world economy, 1960 to 1995. *The Journal of Environment & Development*, *7*(3), 215–247.

Margolis, J. D., & Walsh, J. P. (2003). Misery loves companies: Rethinking social initiatives by business. *Administrative Science Quarterly*, *48*(2), 268–305.

Marti, C. P., Rovira-Val, M. R., & Drescher, L. G. (2015). Are firms that contribute to sustainable development better financially? *Corporate Social Responsibility and Environmental Management*, *22*(5), 305–319.

Matisoff, D. C. (2013). Different rays of sunlight: Understanding information disclosure and carbon transparency. *Energy Policy*, *55*, 579–592.

McGuire, J. B., Sundgren, A., & Schneeweis, T. (1988). Corporate social responsibility and firm financial performance. *Academy of Management Journal*, *31*(4), 854–872.

McWilliams, A., & Siegel, D. (2000). Corporate social responsibility and financial performance: correlation or misspecification? *Strategic Management Journal*, *21*(5), 603–609.

Merton, R. C. (1986). A simple model of capital market equilibrium with incomplete information. *Journal of Finance*, *42*(3), 483–510.

Michelon, G., Rodrigue, M., & Trevisan, E. (2020). The marketization of a social movement: Activists, shareholders and CSR disclosure. *Accounting, Organizations and Society*, *80*, 101074.

Milne, M. J., & Gray, R. (2013). Whither ecology? The triple bottom line, the global reporting initiative, and corporate sustainability reporting. *Journal of Business Ethics*, *118*(1), 13–29.

Mohr, L. A., & Webb, D. J. (2005). The effects of corporate social responsibility and price on consumer responses. *Journal of Consumer Affairs*, *39*(1), 121–147.

Montiel, I., & Delgado-Ceballos, J. (2014). Defining and measuring corporate sustainability: Are we there yet? *Organization & Environment*, *27*(2), 113–139.

Morhardt, J. E., Baird, S., & Freeman, K. (2002). Scoring corporate environmental and sustainability reports using GRI 2000, ISO 14031 and other criteria. *Corporate Social Responsibility and Environmental Management*, *9*(4), 215–233.

Moulton, B. R. (1987). Diagnostics for group effects in regression analysis. *Journal of Business & Economic Statistics*, *5*(2), 275–282.

Muller, A., & Kräussl, R. (2011). Doing good deeds in times of need: A strategic perspective on corporate disaster donations. *Strategic Management Journal*, *32*(9), 911–929.

Murray, A., Sinclair, D., Power, D., & Gray, R. (2006). Do financial markets care about social and environmental disclosure? Further evidence and exploration from the UK. *Accounting, Auditing & Accountability Journal*.

Oberndorfer, U., Schmidt, P., Wagner, M., & Ziegler, A. (2013). Does the stock market value the inclusion in a sustainability stock index? An event study analysis for German firms. *Journal of Environmental Economics and Management*, *66*(3), 497–509.

O'dwyer, B. (2002). Managerial perceptions of corporate social disclosure: An Irish story. *Accounting, Auditing & Accountability Journal*.

Oh, C. H., Park, J.-H., & Ghauri, P. N. (2013). Doing right, investing right:

Socially responsible investing and shareholder activism in the financial sector. *Business Horizons*, *56*(6), 703–714.

Orlitzky, M., Schmidt, F. L., & Rynes, S. L. (2003). Corporate social and financial performance: A meta-analysis. *Organization Studies*, *24*(3), 403–441.

Pablo-Romero, M. d. P., & Sánchez-Braza, A. (2017). Residential energy environmental Kuznets curve in the EU-28. *Energy*, *125*, 44–54.

Perrini, F., & Tencati, A. (2006). Sustainability and stakeholder management: the need for new corporate performance evaluation and reporting systems. *Business Strategy and the Environment*, *15*(5), 296–308.

Pirsch, J., Gupta, S., & Grau, S. L. (2007). A framework for understanding corporate social responsibility programs as a continuum: An exploratory study. *Journal of Business Ethics*, *70*(2), 125–140.

Post, J. E., Preston, L. E., & Sachs, S. (2002). Managing the extended enterprise: The new stakeholder view. *California Management Review*, *45*(1), 6–28.

Qiu, Y., Shaukat, A., & Tharyan, R. (2016). Environmental and social disclosures: Link with corporate financial performance. *The British Accounting Review*, *48*(1), 102–116.

Reinhardt, F. (1999). Market failure and the environmental policies of firms: Economic rationales for "beyond compliance" behavior. *Journal of Industrial Ecology*, *3*(1), 9–21.

Remmen, A. (2007). *Life cycle management: a business guide to sustainability*. UNEP/Earthprint.

Richardson, A. J., & Welker, M. (2001). Social disclosure, financial disclosure and the cost of equity capital. *Accounting, Organizations and Society*, *26*(7-8), 597–616.

Roberts, R. W. (1992). Determinants of corporate social responsibility disclosure: An application of stakeholder theory. *Accounting, Organizations and Society*, *17*(6), 595–612.

Roca, L. C., & Searcy, C. (2012). An analysis of indicators disclosed in corporate sustainability reports. *Journal of Cleaner Production*, *20*(1), 103–118.

Rockstrm, J., Steffen, W., Noone, K., Persson, A., Chapin, F., & Lambin, E. (2009). A safe operation space for humanity. *Nature*, *461*, 472475.

Rodriguez-Fernandez, M. (2016). Social responsibility and financial performance: The role of good corporate governance. *BRQ Business Research Quarterly*, *19*(2), 137–151.

Rosen, C. M., & Sellers, C. C. (1999). The nature of the firm: Towards an ecocultural history of business. *Business History Review*, *73*(4), 577–600.

Rupley, K. H., Brown, D., & Marshall, R. S. (2012). Governance, media and the quality of environmental disclosure. *Journal of Accounting and Public Policy*, *31*(6), 610–640.

Russo, M. V., & Fouts, P. A. (1997). A resource-based perspective on corporate environmental performance and profitability. *Academy of Management Journal*, *40*(3), 534–559.

Saeidi, S. P., Sofian, S., Saeidi, P., Saeidi, S. P., & Saaeidi, S. A. (2015). How does corporate social responsibility contribute to firm financial performance? The mediating role of competitive advantage, reputation, and customer satisfaction. *Journal of Business Research*, *68*(2), 341–350.

Schaltegger, S., Bennett, M., & Burritt, R. (2006). Sustainability accounting and reporting: development, linkages and reflection. An introduction. In *Sustainability Accounting and Reporting* (pp. 1–33). Springer.

Schaltegger, S., Hörisch, J., & Freeman, R. E. (2019). Business cases for sustainability: A stakeholder theory perspective. *Organization & Environment*, *32*(3), 191–212.

Schulz, M., & Kourkoulas, D. (2014). *Regulation (EU) No 423/2014 of the European Parliament and the Council of the European Union.*

Searcy, C., & Buslovich, R. (2014). Corporate perspectives on the development and use of sustainability reports. *Journal of Business Ethics*, *121*(2), 149–169.

Seele, P., & Gatti, L. (2017). Greenwashing revisited: In search of a typology and accusation-based definition incorporating legitimacy strategies. *Business Strategy and the Environment*, *26*(2), 239–252.

Sethi, S. P. (1975). Dimensions of corporate social performance: An analytical framework. *California Management Review*, *17*(3), 58–64.

Shafer, S. M., Smith, H. J., & Linder, J. C. (2005). The power of business models. *Business Horizons*, *48*(3), 199–207.

Shahzad, A. M., & Sharfman, M. P. (2017). Corporate social performance and financial performance: Sample-selection issues. *Business & Society*, *56*(6), 889–918.

Sharfman, M. P., & Fernando, C. S. (2008). Environmental risk management and the cost of capital. *Strategic Management Journal*, *29*(6), 569–592.

Shiu, Y.-M., & Yang, S.-L. (2017). Does engagement in corporate social responsibility provide strategic insurance-like effects? *Strategic Management Journal*, *38*(2), 455–470.

Siano, A., Vollero, A., Conte, F., & Amabile, S. (2017). "More than words": Expanding the taxonomy of greenwashing after the Volkswagen scandal. *Journal of Business Research*, *71*, 27–37.

Singal, M. (2014). The link between firm financial performance and investment in sustainability initiatives. *Cornell Hospitality Quarterly*, *55*(1), 19–30.

Soana, M.-G. (2011). The relationship between corporate social performance and corporate financial performance in the banking sector. *Journal of Business Ethics*, *104*(1), 133–148.

Speelman, D., Heylen, K., & Geeraerts, D. (2018). *Mixed-effects regression models in linguistics*. Springer.

Spicer, B. H. (1978). Investors, corporate social performance and information disclosure: An empirical study. *Accounting Review*, 94–111.

Sroufe, R., & Gopalakrishna-Remani, V. (2019). Management, social sustainability, reputation, and financial performance relationships: An empirical examination of US firms. *Organization & Environment*, *32*(3), 331–362.

StataCorp. (2013). *Introduction to Multilevel Mixed-Effects Models.* StataCorp LP. Retrieved from https://www.stata.com/manuals13/me.pdf

Strike, V. M., Gao, J., & Bansal, P. (2006). Being good while being bad: Social responsibility and the international diversification of us firms. *Journal of International Business Studies*, *37*(6), 850–862.

Surroca, J., Tribó, J. A., & Waddock, S. (2010). Corporate responsibility and financial performance: The role of intangible resources. *Strategic Management Journal*, *31*(5), 463–490.

Székely, F., & Knirsch, M. (2005). Responsible leadership and corporate social responsibility:: Metrics for sustainable performance. *European Management Journal*, *23*(6), 628–647.

Tang, Y., Qian, C., Chen, G., & Shen, R. (2015). How CEO hubris affects corporate social (ir) responsibility. *Strategic Management Journal*, *36*(9), 1338–1357.

Thomson Reuters. (2013). *Thomson Reuters Corporate Responsibility Ratings (TRCRR)*.

Ullmann, A. A. (1985). Data in search of a theory: A critical examination of the relationships among social performance, social disclosure, and economic performance of US firms. *Academy of Management Review*, *10*(3), 540–557.

United Nations. (n.d.). *Sustainable Development Goals.* Retrieved from https://sustainabledevelopment.un.org/sdgs (Retrieved May 6, 2020)

United Nations. (1987). *Report of the World Commission on Environment and Development: Our Common Future.*

United Nations. (2016). *Transforming our world: The 2030 Agenda for sustainable development.*

Van der Laan, G., Van Ees, H., & Van Witteloostuijn, A. (2008). Corporate social and financial performance: An extended stakeholder theory, and empirical test with accounting measures. *Journal of Business Ethics*, *79*(3), 299–310.

Van Marrewijk, M. (2003). Concepts and definitions of CSR and corporate sustainability: Between agency and communion. *Journal of Business Ethics*, *44*(2), 95–105.

Vos, R. O. (2007). Defining sustainability: a conceptual orientation. *Journal of Chemical Technology & Biotechnology: International Research in Process, Environmental & Clean Technology*, *82*(4), 334–339.

Waddock, S. A., & Graves, S. B. (1997). The corporate social performance–financial performance link. *Strategic Management Journal*, *18*(4), 303–319.

Wagner, M. (2007). Integration of environmental management with other managerial functions of the firm: empirical effects on drivers of economic performance. *Long Range Planning*, *40*(6), 611–628.

Wang, L., & Tuttle, B. (2014). Using corporate social responsibility performance to evaluate financial disclosure credibility. *Accounting and*

Business Research, *44*(5), 523–544.
Wang, T., & Bansal, P. (2012). Social responsibility in new ventures: profiting from a long-term orientation. *Strategic Management Journal*, *33*(10), 1135–1153.
WBCSD. (2003). *Sustainable Development Reporting: Striking the Balance.*
Weber, J., & Marley, K. A. (2012). In search of stakeholder salience: Exploring corporate social and sustainability reports. *Business & Society*, *51*(4), 626–649.
Wooldridge, J. M. (2005). Instrumental variables estimation with panel data. *Econometric Theory*, *21*(4), 865–869.
Zellweger, T. M., & Nason, R. S. (2008). A stakeholder perspective on family firm performance. *Family Business Review*, *21*(3), 203–216.