Norwegian School of Economics

Bergen, Spring 2021

Value Relevant ESG Scores and Superior Performance

When is a high ESG score indicative of better financial performance?

Thea Amalie Danielsen and Jan Johansen

Supervisor: Kjell Henry Knivsflå

Department of Finance and Department of Business Analytics

NORWEGIAN SCHOOL OF ECONOMICS

Acknowlegements

The process of writing this thesis has been a rewarding one for both of us, as we were able to spend half a year on a subject that included several of the most interesting subjects from our time as students. The selection of a thesis partner is critical during this process. A partnership based on common strengths is one strategy. Another strategy is to form a partnership based on expertise in distinct but complementary areas. The latter defines the strategy upon which this thesis is based. Ultimately, combining our strengths resulted in a master thesis that we both agree would have been impossible for either of us to complete on our own.

Our supervisor, Kjell Henry Knivsflå, has provided us the capability to communicate our vision and novel ideas in a more effective and professional manner during this process. Whose insightful criticism encouraged us to improve our ideas and raise the bar on our work. We would like to express our gratitude for his help and advice.

We would further like to express our gratitude to our families for their support and wise counsel, their encouragement was crucial in our ability to write this thesis.

Bergen, June 2021

Thea Amalie Danielsen

Jan Johansen

Abstract

Purpose - The objective of this thesis is to develop a methodology that will enable investors to differentiate between two types of companies: those with high performance on ESG issues that have the potential to significantly impact value, and those with high performance on ESG issues that do not have the potential to significantly impact value. **Originality/value** – The Ohlson model is used to calculate a variable based on the value relevance of a consensus ESG score, which allows us to divide companies into two groups: Companies in which ESG has an impact on value and companies in which ESG has no impact on value. **Design/methodology/approach** - To distinguish between companies that allocate resources to ESG issues that are not value relevant and companies that allocate resources to ESG issues that are value relevant, we use the interaction between the variable that divides the companies in our dataset into two groups and the level of the consensus ESG score. This interaction serves as the foundation for the development of two investment strategies, which are tested using two different empirical strategies, one of which represents the creation of actual portfolios applicable to an investor under realistic conditions. **Findings** – We find that an investment strategy based on taking a long position in companies with a high ESG score when ESG is value relevant and a short position in stocks with a high ESG score when ESG is not value relevant generates superior performance, as measured by the Sharpe ratio and the Fama French 5-factor alpha extended with the liquidity factor.

Table of Contents

List of figures

Overview of tables

1. Introduction

The mechanism behind our research question is based on a categorisation of companies with a high ESG score into two groups. The first group mainly allocate their scarce resources to ESG issues that have the potential to significantly impact enterprise value, i.e., peaches. The second group mainly allocate their scarce resources to ESG issues that do not have the potential to significantly impact enterprise value., i.e., lemons. We argue that peaches have a better potential to generate superior performance than lemons. We hypothesize that ESG investors who can efficiently distinguish between these two types of companies would be in a better position to generate superior returns. The main objective of this thesis is to investigate if the value relevance of ESG scores can be used to perform this differentiation. Therefore, the following research question has been formulated:

RQ: Could the value relevance of ESG scores be used to form investment strategies yielding superior performance?

There are two main reasons why our research question is interesting:

The first reason is that this is a timely topic. This is because environmental, social, and governance (ESG) as a concept has grown in popularity among investors, who are increasingly using it to improve the risk-return characteristics of their portfolios rather than simply satisfying the critical societal eye from an ethical standpoint. Sustainable investing in the US increased 25-fold from 1995 to 2020 and has experienced exponential growth since 2012 (US SIF, 2020). In Europe, a similar pattern can be seen (Eurosif, 2016; Eurosif, 2018). The annual net flow to sustainable funds in the United States more than quadrupled between 2018 and 2019, and more than doubled between 2019 and 2020 (Morningstar, 2021). At the start of 2020, the United States had $17.1 trillion in assets under management using sustainable investment strategies, accounting for one-third of the $51.4 trillion in assets under professional management (US SIF, 2020).

The second reason our research question is interesting is that the current most efficient option for performing the described distinction between "lemons" and "peaches" from the perspective of ESG investors is largely ineffective, making research into potential solutions to this challenge valuable. The demand for ESG data has increased in tandem with the exponential growth of ESG investing and the number of investors who use ESG information in their investment analysis. As a result, the market for ESG information intermediaries has expanded. ESG scores are one of the most commonly used forms of ESG data. An ESG score compresses a large body of data into a single score and should reflect how a company is managing its ESG risks and opportunities. The OECD defines ESG scores as investment products (Boffo & Patalano, 2020). Accordingly, they should be able to mitigate information asymmetry such that an investor is able to differentiate between the two types of companies outlined above. However, ESG scores are largely incapable of providing a clear answer on this matter. This is mainly due to the high degree of disagreement between ESG scores from different rating agencies (Semenova & Hassel, 2015; Boffo & Patalano, 2020; Berg, Kölbel, & Rigobon, 2020; Li & Polychronopoulos, 2020; Lopez & Contreras, 2020). The source of the disagreement is to a large degree attributable to the fact that ESG scores are derived from non-financial information which is more ambiguous and much harder to quantify precisely than financial information.

The thesis employs an innovative methodology that combines insights from various strands of the literature. Because of the thesis's objective, it is primarily related to the field of ESG investing. This is due to its goal of developing a method that will allow ESG investors to distinguish between companies with high ESG performance on matters that will benefit an investor holding the stocks, and companies with high ESG performance on matters that will not. Therefore, the thesis's primary contribution is to this part of the literature. The rapid growth of ESG investing has contributed to a large body of literature on ESG investing. As a result, several important papers exist. However, OECD recently released a report that addresses the current state and challenges of ESG investing (Boffo & Patalano, 2020). One of the key challenges highlighted by the OECD is the lack of agreement on what constitutes a financially material ESG issue, i.e., performance on ESG issues that have the potential to benefit investors. Our findings indicate that the method proposed in this thesis has the potential to be an effective tool in assisting ESG investors with the challenge identified by the OECD as the most critical.

The second strand of literature to which we make a significant contribution is one that is closely related to ESG investing and specifically investigates the relationship between ESG performance and corporate financial performance. Friede, Busch, and Bassen (2015) present the most comprehensive study on this relationship, summarizing the findings of over 2000 empirical studies. They find strong evidence for a non-negative link between ESG performance and corporate financial performance. Although the majority of these studies find a positive link, the evidence for a positive link is weaker than that for a non-negative link. Another important article that has influenced our thesis greatly is Khan, Serafeim, and Yoon (2016). The authors of this study found that a link between ESG performance and corporate financial performance is strongest for companies that allocate their resources primarily to ESG issues with the potential to significantly impact enterprise value, which the findings of this thesis validate. Furthermore, Khan, Serafeim, and Yoon (2016) conclude that companies with high ESG performance on ESG issues without the potential to significantly impact enterprise value have no effect on corporate financial performance. Following simple principles, we argue that investing in ESG projects that do not have a positive impact on cash flows or risks represents investments that do not benefit shareholders and, as a result, have a negative impact on value. Our findings provide strong evidence in support of this argument.

Finally, value relevance theory is at the core of the proposed methodology. As a result, our thesis contributes significantly to the literature on the value relevance of ESG scores. The literature on the value relevance of ESG is extensive, but in terms of citations are there no specific papers on this topic that stands out as the most influential. However, the value relevance of ESG has been the primary focus of several research papers, the sum of which is considered influential (De Klerk, de Villiers, & van Staden, 2015; Zuraida, Houqe, & Zijl, 2015; Jain & Rezaee, 2016; Kaspereit & Lopatta, 2016; Miralles-Quirós, Miralles-Quirós, & Valente Gonçalves, 2018). We primarily contribute to this literature in two areas. In the existing literature, it is common practice to use a single ESG score from a single rating agency. As a result of the high level of disagreement among ESG rating agencies, the findings of these research papers are highly dependent on the specific rating agency chosen. By implementing a consensus ESG score derived from seven different ESG scores from six different rating agencies, we contribute to the literature in this area. Furthermore, the literature on the value relevance of ESG suffers from reverse causality bias (Krüger, 2015). Because it is more convenient for profitable companies to devote resources to ESG projects and ESG disclosure in order to improve their ESG score. Since all of the companies in our sample disclose information to all six rating agencies, we have a high disclosure sample. As a result, we eliminate a significant portion of the bias pointed out by Krüger (2015) from our sample. Therefore, we add to the literature by providing a value relevance estimate that is more in line with ESG's incremental effect on value.

Due to the terminology used in our research question we have narrowed the focus to use value relevance theory as a methodological approach to construct ESG investment strategies yielding superior performance. In conjunction with the Ohlson (1995) residual income valuation model, value relevance theory can be used to investigate the ability of non-financial information to explain stock price variation (Amir and Lev, 1996). Value relevance theory is a technique used to investigate how different factors impact company value. The term financially material ESG issues is used in the literature to define ESG issues that have the potential to have a significant impact on a company's value. We claim that a financially material ESG issue is analogous with a value relevant ESG issue. A second important element in the research question is the investment strategies. We use two different investment strategies; one we argue will have the best potential to generate superior performance through a long position and the other using a short position. To determine which companies that allocate resources to ESG issues that can explain the stock price, we use the residual difference between the Ohlson (1995) model with and without the consensus ESG score. Using this variable in conjunction with the level of the consensus ESG score, we can divide companies into two groups, as described above, ref. lemons and peaches. Because of the aforementioned relationships, we claim that the lemons group is the best option for an ESG investment strategy based on a short position. Following the same logic, we argue that the peaches group is the best option for an ESG investment strategy based on a long position. We find that an investment strategy that takes a short position in the peaches group and a long position in the lemons group generates superior performance. As a result, our findings provide strong evidence that validates the claimed relationships.

The structure of the thesis is as follows. Section 2 provides an overview of the key components of our research question: ESG, superior performance, and value relevance. The 3rd section contains the hypothesis development and testing methodology. Section 4 provides details on our data, variable construction, descriptive statistics, and simple correlations. Section 5 describes our robustness tests' methodological approach. The findings are presented in section 6. Our conclusion is found in Section 7.

2. ESG, Superior Performance and Value Relevance

2.1 Environmental, Social and Governance (ESG)

Environmental, social and governance (ESG) is one of several frameworks available for identifying a company's sustainability practices. At the moment, ESG is the most popular. Corporate Social Responsibility (CSR) is another popular framework, but its use has declined as the importance of environmental issues has grown. This is because CSR effectively is synonymous with ESG without the environmental pillar. Several of principles used to define the ESG criteria have evolved from Corporate Social Responsibility (CSR) (Boffo & Patalano, 2020). Therefore, are the findings in CSR research to a large extent generalizable to ESG. Furthermore, many studies choose not to use the ESG terminology and instead use the more general term: sustainability, despite having the same focus as studies that use the ESG terminology.

ESG is based on three pillars: environmental, social, and governance. A company that performs well on the environmental pillar operates in such a way that environmental harm is minimized. Climate change and carbon emissions, pollution, resource depletion, waste generation, and deforestation are all part of this pillar. Climate change and carbon emissions are the most important specific ESG issues that money managers consider within the environmental pillar (US SIF, 2020). Business relationships, customer relationships, human rights, working conditions, child labour, and employee relations are examples of issues addressed in the social pillar. Conflict risk is the most important specific social pillar issue for money managers and considered the overall most important ESG criterion for institutional investors (US SIF, 2020). Finally, the governance pillar refers to how an organization is run and is frequently regarded as the most important by investors because poor governance may imply that ESG criteria in general are poorly managed (Boffo & Patalano, 2020). The governance pillar is concerned with issues such as information transparency, tax strategy, executive pay, bribery and corruption, and board diversity and structure. Anti-corruption is considered the most important specific governance issue, according to money managers (US SIF, 2020). Despite the fact that disclosure of information in accordance with ESG-criteria is voluntary, the number of companies measuring and reporting their ESG practices has grown exponentially in recent years. This is due to its importance as a tool in investor relations management and public communication in general as part of a branding strategy.

Since a large part of the terminology used in ESG has evolved from corporate sustainable responsibility (CSR), there is a lack of standardization in the vocabulary used to describe ESG investing. As a result, it can be difficult to tell the difference between ESG investing and sustainable investing, also known as socially responsible investing (SRI) or impact investing. S&P Global (2020) attempts to clarify the ambiguity by defining ESG investing as taking into account ESG risks and opportunities when making investment decisions. Additionally, they actively incorporate this into their investment analysis in order to invest sustainably while not sacrificing returns and, ideally, increasing profits. Sustainable investing, on the other hand, prioritizes social change over capital gains by taking both financial and moral values into account when making investment decisions.

2.1.1 ESG scores

ESG scores compress information from a large body of data into a single score typically on a scale from 0 to 100, with 100 representing an excellent score and 0 representing a poor score. ESG rating agencies are classified as information intermediaries. Hence, one of their key functions is to reduce information asymmetry (Healy & Palepu, 2001). Using a framework developed by Li and Polychronopoulos (2020), ESG rating providers can be classified into three categories. The first, *fundamental*, consists primarily of ESG data providers. Their primary function is to collect and aggregate publicly available data; they typically do not provide an overall ESG score for companies. The most common category is the second, *comprehensive*. In the year 2020, there were about 70 providers offering ESG rating data, with 50 of them falling into the comprehensive category (Li & Polychronopoulos, 2020). While ESG data vendors in the fundamental category only use objective data, ESG rating agencies in the comprehensive category use far more subjective data. That is, data that they have collected, or data generated by their own analysts. As will be discussed in the subsection addressing ESG score disagreement, this is a source of measurement divergence across rating agencies. Finally, there is the specialist category, which focuses in depth on a specific criterion. For example, an ESG score that focuses solely on carbon emissions would fall into this category.

Most rating agencies prioritize two factors in their ESG scores: ESG risks and opportunities that have the potential to have a significant impact on the company's value. These ESG issues are defined as material ESG issues in the literature. The rating agencies consider how a specific company is exposed to material ESG issues relative to industry peers, as well as how they manage these risks and opportunities. ESG scores are investment products by definition, but in order to be relevant to investors, the metrics used by rating agencies to generate ESG scores must reflect underlying ESG issues that have the potential to significantly impact enterprise value, i.e., material ESG issues (Boffo & Patalano, 2020). The SASB is a US-based non-profit organization that develops guidelines for businesses to disclose material sustainability information to investors (SASB, 2021). According to SASB, material ESG issues are those that are most likely to have a direct impact on a company's financial condition or operating performance, and thus are most important to investors (SASB, 2021). Rating agencies typically emphasize that their ESG scores focus strictly on financially material ESG issues. There appears to be agreement amongst the rating agencies that material ESG issues are those that have the potential to significantly impact a company's value, either negatively or positively (Refinitiv, 2021) (Sustainalytics, 2020) (MSCI, 2020). Sensitive industries are another concept that is similar to the term materiality; sensitive industries will be considered industries with a high concentration of material ESG issues when applying the materiality term to that context. Companies in environmentally sensitive industries are more likely to face potential corporate social responsibility disclosure litigation and future environmental liability, as well as increased public attention and concern (De Villers, Naiker, & Van Staden, 2011). As a result of these risks, investors in environmentally sensitive industries would have more cause to be worried about negative cash flow effects from environmental and other social issues. Concerns that could be significantly mitigated if the company has a high ESG score. However, if the ESG score does not reflect performance on material ESG issues, it will not be able to reflect the true risk exposure.

In an effort to use consistent and precise language, we will replace the term *material* with the term *significant* where appropriate. This is because a material ESG issue is defined as an ESG issue that has the potential to significantly impact a company's enterprise value. Based on this, we believe the interpretation is nearly identical.

ESG rating agencies

In this thesis, we use seven different ESG scores from six different rating agencies. Five of these scores are in the comprehensive category: Refinitiv Eikon ESG, Refinitiv Eikon ESG-C, Sustainalytics ESG Risk Ratings, RobecoSAM Total Sustainability Score, and Bloomberg Disclosure score. All these five scores are calculated as a weighted sum of the E, S, and G pillars, with weights individually determined by the rating agency. The ISS Quality Score and the Carbon Disclosure Project's (CDP) Climate, Water, and Forest score both fall into the specialist category. The final scores from these two agencies are not based on the three E, S, and G pillars.

Refinitiv Eikon ESG Score (Refinitiv, 2021). The Refinitiv Eikon ESG scores – previously Thomson Reuters – incorporates a company`s ESG performance, commitment, and effectivity. An interesting feature of the Refinitiv ESG scores is that they use a materiality matrix that through the weightings of the 10 categories, behind the three pillars, implement materiality into the scores, the materiality weightings is industry specific. Several investors rated the Refinitiv Eikon ESG scores as low quality, while none rated them as high quality. However, several of the survey's expert participants rate Refinitiv Eikon as high quality (SustainAbility, 2020). The raw data, rather than the scores themselves, is cited by investors as the primary reason for using the ratings (SustainAbility, 2020).

Refinitiv Eikon ESG-Controversies Score (Refinitiv, 2021). Refinitiv also provides an overall ESG-C score, which penalizes a company's standard Refinitiv ESG score if it is involved in major ESG controversies that have been covered by the global media. Large-cap companies are more vulnerable to controversies because they receive more media attention than smaller-cap companies. As a result, they receive a higher severity weight, resulting in a higher penalty to the ESG score. The ESG controversies ranking is based on 23 different ESG controversies topics that are updated on a regular basis.

RobecoSAM Total Sustainability Score (RobecoSAM, 2021). The key data source used to generate the RobecoSAM ESG ratings is the Corporate Sustainability Assessment, an annual survey of over 4,700 companies that focuses on ESG-related issues that are industry specific and material. Constructing a portfolio of companies with high ESG scores can often lead to exposure to factor biases such as size and quality, or it can result in a portfolio that is disproportionately overconcentrated in some countries or industries. RobecoSAM claims that their ESG score, which they refer to as a smart ESG score, is free of these biases. RobecoSAM is rated high quality by most investors and second to highest by the experts in the same survey (SustainAbility, 2020). However, investors claim that the score is not used for investment decisions because the methodology is opaque (SustainAbility, 2020).

Sustainalytics ESG Risk Ratings (Sustainalytics, 2019). While other ESG ratings serve as indicators of multiple parameters, Sustainalytics scores only concentrate on ESG risks. Specifically, the unmanaged risks, which consist of two parts: risks that cannot be managed by the company and risks that could be managed but is not. Only significant ESG risks is included, implying only risk exposure that would leave the economic value of the company at risk. Sustainalytics use corporate governance in combination with exposure to significant ESG issues as important building blocks to arrive at the final score. Sustainalytics is one of the most frequently mentioned ratings by investors, with the preference for Sustainalytics being largely driven by its broad coverage, according to investors in a survey (SustainAbility, 2020).

Bloomberg Disclosure Score (Bloomberg, 2021). The level of a company's environmental, social, and governance (ESG) disclosure at each of the datapoints collected by Bloomberg determines the Bloomberg Disclosure Score. Therefore, the scores do not take ESG performance into account. The ranking ranges from 0.1 for businesses reporting the bare minimum of ESG data to 100 for those disclosing every data point. Each data point is weighted in terms of importance, which is customized to different sectors due to substantial industry specific differences. This ensures that companies are only evaluated on data points that important in each industry. The Bloomberg ESG disclosure score is rated high and low quality by the same number of respondents in the Rate the Raters survey, but few investors mention using this score (SustainAbility, 2020).

CDP Climate, Water, Forest score (CDP, 2021). This specialist ESG score is essentially an ESG score with only the environmental pillar. CDP`s ESG scores promotes organizational accountability while also assisting in the direction, incentivizing, as well as evaluating environmental action. Climate change, deforestation, and water security are the three key pillars that make up the CDP performance score. In the Rate the Raters, investors stated that they use CDP mostly for its data, it is rated almost highest by investors and highest by the experts in the survey (SustainAbility, 2020).

ISS Quality Score (ISS, 2021). This ESG score is mostly concerned with the government pillar. In the areas of Board Structure, Compensation Plans, Shareholder Rights, and Audit & Risk Oversight, the ISS QualityScore score is intended to assist institutional investors in measuring quality benchmarks and assessing risk. The score can explain how a company's approach toward governance has changed over time. Several investors in the Rate the Raters survey stated that ISS had the best governance report (SustainAbility, 2020).

ESG score disagreement

Both ESG scores and credit scores provide a single datapoint representation derived from the analysis of multiple datapoints. However, the two scores' reliability and validity deviate significantly. Credit scores from a variety of rating agencies have an average correlation of 0.986 (Lopez & Contreras, 2020), whereas the average correlation between the six possibly most widely used ESG rating agencies was found to be only 0.46. (Gibson, Krueger, & Schmidt, 2020). The primary reason for the low correlation across rating agencies is that financial information leaves less room for subjective interpretation than non-financial information.

Berg, Kölbel, and Rigobon (2020) developed a clear theoretical framework for formally quantifying the source of disagreement, tracing the cause of the divergence to three distinct sources. First, they define *scope divergence* as the attributes that are included in the score, thereby determining what the ESG score can measure. Second, *measurement divergence* refers to the fact that even when rating agencies agree on the inclusion of a variable, they may arrive at different conclusions even though they want to measure the same thing. Non-financial information requires a wide use of proxies, and measure divergence could for instance be attributable to the use of different proxies. Measurement divergence is found to be the biggest driver of the divergence across ESG scores (Berg, Kölbel, & Rigobon, 2020). Finally, *weight divergence* occurs as a result of ESG scores compressing a large number of data points into a single score. As a result, rating agencies will differ in how they aggregate the data points into a single score. Divergence in ESG scores due to scope divergence is positive because it leads to the scores reflecting slightly different aspects of sustainability performance; however, divergence in ESG scores due to measurement divergence is negative (Berg, Kölbel, & Rigobon, 2020).

The most recent Rate the Raters Survey (SustainAbility, 2020) revealed several interesting aspects about how professional investors use ESG scores. Investors appear to be aware of the significant inconsistency in ESG scores. Because almost every respondent stated that they never rely on a single ESG rating during their investment research. Furthermore, investors do not rely solely on ESG ratings in their investment research; rather, they view them as one data point in a larger picture. However, the low correlation across ESG rating agencies suggests that the quality of the ESG ratings also will vary significantly. This gives investors a strong incentive to find the rating agency with the best ESG scores. The degree to which an ESG score reflects significant ESG issues is one of the top three factors used to rank rating agencies relative to one another. When asked about preferred changes to ESG scores in the next five years, the majority stated that they wanted ESG scores that reflect ESG issues that have the potential to have a significant impact on the enterprise value, i.e., significant ESG issues.

2.2 Superior Performance

In this section, we will provide an overview of the factors that must be present to generate superior performance, as well as whether there is a plausible reason to believe that ESG can contribute to it.

2.2.1 The link between ESG performance and superior performance

For a long position in ESG stocks to produce superior performance, there need to be a positive link between ESG performance and superior performance. There are several ways to explain how ESG performance can lead to outperformance. As a foundation to do this, we chose a model used in a research article by Giese et al. (2019-a). This is because their study on the relationship between ESG performance and superior performance is based on the well-known and traditional discounted cash flow (DCF) model. Therefore, Giese et al. (2019-a) provide an easy-to-understand framework for describing the transmission channels that can generate a causal link between ESG performance and possible superior performance. This is accomplished using a methodology in which the authors investigate whether a change in a company's ESG performance affects the respective financial target variables, implying improved corporate financial performance and, as a result, a higher valuation. In the DCF model, ESG performance has only two potential channels for influencing the company's value: 1. The systematic channel that will act through the fraction's denominator, reflecting systematic risk via the cost of capital. 2. The idiosyncratic channel acting through the fraction’s numerator, which reflects profitability and idiosyncratic risk via cash flows. Giese et al. (2019-a) find that ESG performance link to outperformance is transmitted through both channels. ESG performance was discovered to be transmitted through the idiosyncratic channel, affecting cash flows in such a way that it resulted in higher profitability and lower tail risk. The increased profitability of ESG companies was attributed to a competitive advantage in areas such as efficient resource utilization, human capital development, innovation, and being better at designing long-term business strategies, making them less susceptible to the pitfalls of short-termism. ESG performance was also transmitted via the systematic channel, as it resulted in changes in the financial target variables in the DCF model's denominator, resulting in lower cost of capital, which led to higher valuations. This was attributed to the notion that companies with high ESG performance are less susceptible to market shocks. The authors point out that in a portfolio context where the idiosyncratic transmission channel is diversified away, the investor will only gain from the benefits transmitted through the systematic risk channel.

Even though several studies have identified ways in which ESG performance can be linked to corporate financial performance, construct validity remains a major challenge in this research. This is due to a widespread reverse causality bias, which makes it difficult to attribute a positive relationship between ESG performance and superior performance to a causal effect. According to Krüger (2015), research that investigate the correlation between superior performance and high ESG performance are unable to answer the fundamental question of whether firms are more profitable because they have a high ESG rating or are more profitable because they initially were more profitable. As a result, a finding of a positive correlation between a high ESG score and value can be due to either the fact that businesses with high ESG performance are more profitable or, instead, that companies with high ESG performance allocate more resources into projects and ESG disclosure to enhance their ESG score. As a result, if the relationship between ESG and superior performance is not causal, the potential gains from investing in top-rated ESG stocks may be less consistent over time.

2.2.2 Significant ESG issus have a stronger link to superior performance

From the standpoint of agency theory, one might wonder if a board of directors pursuing ESG is acting in the best interests of their shareholders. According to Milton Friedman and shareholder theory, a corporation's aim is to be profitable, and it should not spend its resources or cash flows on social responsibility other than adhering to society's basic laws (Friedman, 1970). However, with key stakeholders increasingly emphasizing ESG factors, ignoring them can have a negative impact on profits. Edvard Freeman with his stakeholder theory question shareholder theory and its narrow focus on shareholder benefit, arguing that companies should consider all their stakeholders' interests (Freeman, 2010). However, focusing on all stakeholders while maintaining profitability will be difficult. 50 years after Friedman's shareholder theory, Jay Barney, one of the most influential figures in the strategy field, proposes that companies only should consider the interests of stakeholders who have an influence on the company's profit (Barney, 2020). When applied to ESG, Barney is essentially arguing that the board of directors should only focus on ESG issues that have the potential to have a significant impact on enterprise value, an argument that even Friedman might agree with.

As previously stated, we will replace the term *material* with the term *significant* where appropriate in order to use precise and consistent language. However, it is appropriate to return to the use of the term materiality in this brief segment. Khan, Serafeim, and Yoon (2016) wrote a highly influential article that to a large extent have influenced the widespread use of the term materiality; their findings are in many ways consistent with Barney (2020). In simple terms, their findings show that a company that allocates its resources primarily to ESG or sustainability issues that have the potential to significantly impact enterprise value increases the likelihood of ESG performance resulting in superior financial performance.

However, almost all rating agencies have developed their own methods of defining what constitutes material ESG issues within each industry. Additionally, several organizations that offer sustainability reporting standards also provide definitions of significant/material ESG issues. The Global Reporting Initiative (GRI), the Task Force on Climate-related Financial Disclosures (TCFD), the International Integrated Reporting Council (IIRC), and perhaps the most influential, the Sustainability Accounting Standards Board (SASB), are the most well-known of these organizations. As a result, while there is agreement that significant ESG issues result in a stronger link between ESG performance and corporate financial performance, there is little consensus on which ESG issues should be defined as material. One of the most important differences is closely linked to stakeholder and shareholder theory. GRI, for example, employs a concept of materiality with all stakeholders as their target audience, which is analogous to stakeholder theory. Unlike GRI, SASB uses the US Supreme Court definition of materiality, which has investors as the clear target audience rather than all stakeholders, which is more consistent with shareholder theory.

As previously stated, there appears to be an emerging consensus among ESG rating agencies regarding the use of the SASB definition of material ESG issues; however, the agencies are still developing their own materiality maps. Hence, while there will be some agreement between rating agencies on which ESG issues to define as material within each industry, which is highly influential in determining the ESG score for companies operating in these industries, there will still be several ESG issues on which rating agencies disagree. As a result, users of ESG scores will be less confident that the ESG scores accurately reflect the ESG issues that have the greatest potential to influence the stock price of their investments.

2.2.3 Purchasing high ESG performance witout having to pay for it upfront

Even under the assumption that there is a significant relationship between ESG performance and superior performance, will this not imply that constructing a portfolio of companies with high ESG performance will result in superior performance. Fama (1970) classified market efficiency into three levels: weak, semi-strong, and strong. Empirical evidence from several studies presented in Fama (1991) and Fama (1998) demonstrates that the market is efficient at the semi-strong level. When the market is efficient at the semi-strong level it implies that prices reflect all publicly available information. As a result, companies that perform well on significant ESG issues will charge a premium unlike companies that perform poorly on these issues. Hence, paying a premium for the predicted future superior performance of stocks with high ESG performance will not produce superior returns for a potential investor. As a result, if markets are effective at a semi-strong level, we will be unable to develop an investment strategy that generates superior and consistent long-term performance based on a long position in ESG stocks.

There are however a few known deviations from efficient market hypothesis (EMH), for instance mean reversion as first discovered by DeBondt and Thaler (1985), who found that winners and losers in one 36-month cycle appear to reverse their results in the next 36-month period. Mean reversion actually supports a view that developing an investment strategy based on a short position in high-ranked ESG stocks can generate superior returns. This is because part of the exponentially increasing attention ESG stocks have received in recent years has been due to their status as winners. However, assuming long-term negative serial correlation as predicted by mean reversion, the end result could be paying for superior future performance that does not materialize. Another deviation is that there appears to be positive serial correlation in the short term, as stocks that rise tend to rise further (Jegadeesh & Titman, 2001). Deviations between price and value caused by momentum can be exploited through short-term trading strategies. Exploiting short-term deviations from value, on the other hand, carries a high level of risk, and as we will see in the following section, this is not a strategy used by institutional ESG investors, making this deviation of little use in generating superior consistent returns from ESG stocks. Additionally, it is regarded as a proxy for risk (Jegadeesh & Titman, 2001).

Consequently, we conclude that markets are, for the most part, effective. This implies that, in general, ESG performance as reflected through ESG scores only, as a rule, will be able to generate superior returns if the ESG scores are used in a way that differs from what other market participants do in aggregate. This would be related to what is referred to as "*opportunities in secondary or little-known issues*" in the chapter on discrepancies between price and value in the classic book on security analysis by Graham and Dodd (2009). This would be required in order for you to be able to purchase ESG stocks that will outperform in the future but without paying for it in advance.

2.2.4 Common ESG investing strategies

OECD (2017) have identified six key investment strategies that is used by institutional investors to incorporate ESG considerations into their portfolio construction process. The most common method of ESG investing is *exclusionary screening*. This typically implies excluding a sector, country or a company based on one or several ESG features. The benefit of exclusionary screening is that it can be implemented cheaply and easily. Roughly 44% of professionally managed assets in Europe follow exclusionary screening to some degree (Eurosif, 2016), but this approach has experienced a decline in recent years (Eurosif, 2018). *General ESG integration* is a technique that incorporates ESG risks and opportunities into conventional investment research with the objective of improving the risk-return characteristics of the portfolio. This ESG investment strategy has grown in popularity and experienced rapid growth in recent years, with a CAGR of 27% between 2016 and 2018. Another ESG investment technique that is gaining traction is the *best-in-class* approach. Best-in-class has grown at a healthy 20 percent CAGR in Europe over the last eight years, reaching nearly 600 billion euros, solidifying its status as one of the top ESG investment strategies (Eurosif, 2018). For investors following the best-in-class screening methodology no sector is excluded, and ESG scores are typically applied to include companies within a sector that rank higher than a pre-set hurdle (OECD, 2017). Then there is the *thematic investment strategy*, which chooses an investment universe based on an ESG-related theme. The opposite of the thematic approach, the *de-investment strategy* involves opting out of a particular investment universe based on ESG-related themes. The last ESG investment strategy defined by OECD (2017) is the *engagement strategy*, which implies using ownership to shift the focus of a company towards ESG. This interaction strategy is already a mainstream strategy, and it is improving its position in comparison to other strategies, with a CAGR of 7% from 2016 to 2018. As a result, it seems that investors are increasingly exercising their ownership rights to shift their companies' commitment to sustainability (Eurosif, 2018). This investment strategy prioritizes social change in investment decisions, considering both financial and moral values, with capital gains as a secondary consideration (S&P Global, 2020).

To summarize, there appears to be a trend in which institutional investors are shifting their ESG investing strategies away from those with an ethical emphasis and toward those with the best prerequisites for using ESG to generate superior performance and enhance the risk-return characteristics of portfolios. As evidenced by the significant increase in investment strategies such as general ESG integration and the best-in-class approach, while exclusionary screening is declining.

2.3 Value Relevance

In simple terms, value relevance theory is a technique used to investigate how various factors influence company value. Value relevance is defined as *“the ability of financial statement information to capture and summarise information that determines the firm`s value”* (Beisland, 2009). The core issues studied in this line of research is to which extent various sources of financial information ex-post can explain price fluctuations. However, since Amir and Lev (1996), using the Ohlson (Ohlson, 1995) residual income valuation model to also investigate the value relevance of non-financial information has become a common approach. The Ohlson model derives the market value of equity from two primary sources of information: the company's financial information and value-relevant non-financial information on which Ohlson did not elaborate (Ohlson, 1995). The financial information is specified as the book value of equity and abnormal earnings. The abnormal earnings are in the model defined as the residual income; hence it is often called the Ohlson residual income valuation model. Because of the lack of consensus in estimating the cost of capital needed to arrive at the residual income, a proxy is commonly used in its place. The Ohlson model as defined by Collins, Maydew, and Weiss (1997) use earnings per share to proxy residual income and is defined on a per share basis. The per share version of the Ohlson model was later recommended (Barth & Clinch, 2009). Equation 1 shows the Collins, Maydew, and Weiss (1997) version of the Ohlson model:

$$Stock\ Price_{i,t} = \beta_0 + \beta_1 BVPS_{i,t,} + \beta_2 EPS_{i,t} + \beta_3 OI_{i,t} + e_{i,t} \quad (1)$$

The subscripts i and t represent the company and fiscal year, respectively. Beta 0 to Beta 3 represent the parameters that will be estimated; e is the residual. The independent variable is the stock price, which is represented by the last closing price for a given company on the last trading day of the year. BVPS stands for book value per share as of the end of the fiscal year; EPS stands for earnings per share as of the end of the fiscal year; and OI stands for other value relevant non-financial information.

A company's stock price is expected to reflect all available information (Fama, 1991). Accordingly, the price of a stock, as reflected in the Ohlson model, is also influenced by factors other than financial information. To the best of our knowledge, Amir and Lev (1996) were the first to use the Ohlson model to examine the value relevance of non-financial information as the content of the Ohlson model`s additional unspecified third dependent variable. This approach has since been adopted in several studies investigating the value relevance of company’s sustainability practices (De Klerk, de Villiers, & van Staden, 2015; Zuraida, Houqe, & Zijl, 2015; Kaspereit & Lopatta, 2016; Miralles-Quirós, Miralles-Quirós, & Valente Gonçalves, 2018). Equation 2 shows the Collins, Maydew, and Weiss (1997) version of the Ohlson model when used to investigate the value relevance of ESG scores.

$$Stock\ Price_{i,t} = \beta_0 + \beta_1 BVPS_{i,t,} + \beta_2 EPS_{i,t} + \beta_3 ESG_{i,t} + e_{i,t} \quad (2)$$

With the addition of the ESG score variable for company i at year t. The ESG score variable represents non-financial information that is believed to be value relevant.

3. Hypothesis development and test methodology

In this section, we will describe the hypothesis development and testing methodology used. The research question in this thesis is formulated as follows:

Could the value relevance of ESG scores be used to form investment strategies yielding superior performance?

There are three variables in our research question: value relevance to ESG scores, investment strategies, and superior performance. The *value relevance of ESG scores* is the independent variable which is in a causal relationship with the dependent variable *superior performance*, implying that the value of the independent variable influences the value of the dependent variable. The investment strategy influences the effect of the independent variable on the dependent variable. Accordingly, it is an interaction variable. Our findings can be simplified to result in each of the three variables taking two main values: The independent variable can take the values: 1. *value relevant*, or 2. *not value relevant*. We will diversify between two investment strategies, such that the interaction term can take the values: 1. *short position*, or 2. *long position*. The values of the independent variable and the interaction variable can cause the dependent variable to take the values: 1. superior performance, or 2. non-superior performance. Superior performance is not a standardized term, so it must be operationalized. We will use two metrics to evaluate if an investment strategy can generate superior performance. First, we will look at whether the investment strategy improves the risk-return relationship as measured by the Sharpe ratio. Second, we will use the five-factor alpha (Fama & French, 2015) combined with the liquidity factor (Pástor & Stambaugh, 2003) as the ultimate measure of whether the investment strategy can generate superior performance.

3.1 Hypothesis development

In developing our hypothesises to answer our research question we will first focus on the value of the independent variable: *value relevance of ESG scores*. This is because we need to determine whether ESG scores are value relevant to test whether value relevant ESG scores will allow us to create investment strategies that will generate superior performance. This will be accomplished by reviewing existing literature on the value relevance of ESG. Next, we will focus on the dependent variable through previous research covering adjacent topics, as no previous research have covered the specific link between value relevant ESG scores and superior performance. Starting with the link between ESG performance and superior performance before moving on to other studies that have attempted to use ESG scores as a source of alpha generation. In terms of the interaction variable, we will choose one investment strategy based on a short position and one based on a long position. Established on an overall assessment of what has been covered to maximize the possibility of the strategy generating superior performance. The overall objective is validating that value relevant ESG scores can be used to create investment strategies yielding superior performance.

3.1.1 The value relevance of ESG

The first two value relevance studies examined are unrelated to ESG, but they are among the most frequently cited value relevance studies. The first paper by Collins, Maydew, and Weiss (1997) is included to demonstrate the consistency of the Ohlson model using only earnings and book values, while Amir and Lev (1996) is included because it was the first paper to use the Ohlson (1995) model to study the value relevance of non-financial information.

As the millennium approached, tech companies were on the rise, leading to what became known as the dot.com bubble. Businesses with a high concentration of intangible assets were becoming a larger part of the economy. Several research papers argued that as a result of this shift, the value relevance of earnings and book values had deteriorated. Collins, Maydew, and Weiss (1997) investigated whether the value relevance of earnings and book values has shifted over the last four decades in this context. Their findings suggest that earnings are less value relevant because there is a higher concentration of intangible-intensive companies and companies with negative earnings. Despite their findings that the value relevance of bottom-line earnings has declined, the value relevance of book value has picked up the slack and become more value relevant. As a result, the combined value relevance of earnings and book values has remained stable. Hence, earnings and book values combined are consistently value relevant. Amir and Lev (1996) use the Ohlson model to study the value relevance of non-financial indicators. This on a sample of companies in a fast-changing, high-growth technology industry, as a large portion of these companies' value is derived from intangibles, the authors hypothesized that traditional financial indicators would be less value relevant. According to the findings, the value relevance of non-financial information overwhelms that of traditional financial indicators. Furthermore, they discover that when earnings are combined with non-financial information, the value relevance of earnings increases.

Hassel, Nilsson, and Nyquist (2005) applied the same methodology as Amir and Lev (1996), but with a specialist ESG score only focusing on the environmental pillar as the non-financial information. Hassel, Nilsson, and Nyquist (2005) use an empirical analogue of the Ohlson model to explain how environmental information is reflected in the market value of Swedish companies. However, because this study was conducted prior to the massive increase in ESG data availability that we have seen over the last decade, environmental performance is proxied by an index designed specifically for institutional investors. Their findings indicate a significant negative relationship between the market value of Swedish companies and the environmental performance index score; the findings are also applicable across industries. We argue that the use of a potentially low-tech environmental score could explain the negative value relevance discovered by Hassel, Nilsson, and Nyquist (2005), as it is not unlikely that this score fails to distinguish between significant and non-significant ESG issues.

Another study that finds the value relevance of ESG to be negative is Landau et al. (2020). They look at how integrated reporting of ESG and financial data affects the market valuation of 50 European stocks from 2010 to 2016. The findings show that this has a negative impact on market valuation. The authors used the same empirical analogue of the Ohlson model as Hassel, Nilsson, and Nyquist (2005), but in a natural logarithm format to allow for the wide range between variables. We argue that the negative relationship discovered by Landau et al. (2020) could possibly be attributed to sample bias because of a sample size of only 50 companies. As for instance Aureli et al. (2019) in a similar study, but with a bigger sample reached the opposite conclusion. Aureli et al. (2019) study the value relevance of ESG disclosure on Dow Jones Sustainability World Index companies between 2009 and 2016. They use an event study as a methodological approach to investigate how investors react to ESG information released on company websites. The findings indicate that the value relevance of this information is positive and significant, and that it has increased since 2013.

Semenova, Hassel, and Nilsson (2010) find environmental and social information to be value relevant. They use the same empirical analogue of the Ohlson model as Hassel, Nilsson, and Nyquist (2005) to investigate how environmental and social information is reflected in the market value of Swedish firms. As a proxy for environmental and social data, the GIS Investment Services Risk Rating database is used. The results indicate that companies with better environmental and social performance outperform the market, while those with worse performance underperform. Zuraida, Houqe, and Zijl (2015) find the Bloomberg disclosure ESG scores to be value relevant. They use the Collins, Maydew, and Weiss (1997) version of the Ohlson model to analyse the value relevance of ESG disclosure on an international sample of 38 countries between 2008-2012. The Bloomberg disclosure scores are used as a proxy for ESG disclosure, as pointed out by the authors have this approximation not been checked against other ESG-metric providers. The findings show that the disclosure of ESG information has a positive impact on the valuation of companies. Their regressions on the individual E, S and G pillars show that the social pillar has the lowest positive association with stock price, while governance has the strongest. De Klerk, de Villiers, and van Staden (2015) use the Barth and Clinch (2009) version of the Ohlson model on a sample of the 100 largest UK companies to study the value relevance of corporate social responsibility (CSR) disclosure. They find that CSR disclosure is value relevant and that it has a positive effect on the market value, this effect is stronger for companies operating in sensitive industries. This study stands out as particularly robust as they separately estimate the Ohlson model using three different proxies for CSR. Jain and Rezaee (2016) investigate whether short sellers use ESG information to make investment decisions. They discover a negative relationship between rising ESG scores and short selling, and the authors conclude that this is because ESG scores are value relevant. Jain and Rezaee (2016) also conduct a value relevance study, using the Collins, Maydew, and Weiss (1997) version of the Ohlson model to test the value relevance of the Bloomberg Composite ESG Score and the KLD Composite ESG Score. According to the regression results, both ESG scores are value relevant, with KLD being slightly more so. However, because the companies in the KLD and Bloomberg datasets are not identical, the difference could be due to sample variation. The sample period runs from 2004 to 2012. Kaspereit and Lopatta (2016) investigated the value relevance of corporate sustainability in the 600 largest European companies from 2001 to 2011. The SAM sustainability ranking, and sustainability reporting are used as a proxy for corporate sustainability and the Feltham and Ohlson valuation model is the empirical model. According to the findings, the SAM sustainability ranking is value relevant. Suggesting a positive relationship between market value and corporate sustainability.

Even though not all the eight papers reviewed specifically used the ESG terminology, they do cover aspects that would fall under the definition of ESG issues. Six of the papers are directly relevant to determining whether the independent variable in our research question is likely to take the value, value relevant, because these used some form of ESG score as a proxy for underlying ESG performance. The fact that five out of six studies that used some form of ESG scores found the scores to be value relevant increases the likelihood that the ESG scores in our sample will be value relevant as well. However, as discussed in section 2, studies attempting to establish a link between ESG performance and abnormal performance are plagued by reverse causality bias, as Krüger (2015) points out. This is because it is easier for profitable companies to allocate resources to ESG disclosure and ESG projects, which will generate a higher ESG score. Companies in our sample that do not provide information to all six rating agencies are removed from the dataset, resulting in a sample with a high level of disclosure. In terms of causality, this is advantageous because our estimate of value relevance will be closer to the causal effect of ESG on company value. However, because of this characteristic in our sample, the ESG scores are probably less value relevant than in the reviewed studies, which found ESG scores to be value relevant regardless of whether they were from GSI (Semenova, Hassel, & Nilsson, 2010), GRI (De Klerk, de Villiers, & van Staden, 2015), Bloomberg (Zuraida, Houqe, & Zijl, 2015; Jain & Rezaee, 2016), KLD (De Klerk, de Villiers, & van Staden, 2015), or SAM (Kaspereit & Lopatta, 2016). Although two of the value relevance studies, we reviewed used more than one ESG score they did not have a high disclosure sample. De Klerk, de Villiers, and van Staden, (2015) used three proxies for CSR performance, but all these scores were from GRI. Jain and Rezaee (2016) used ESG scores from both Bloomberg and KLD, but they used different datasets for each score. As a result, did not companies have to disclose information to both rating agencies to avoid being dropped from their datasets. Therefore, the reviewed studies cannot be considered to have been conducted on a high-disclosure sample. As previously stated, a large portion of the bias identified by Krüger (2015) is not a major concern in our study due to our high disclosure sample. However, our high disclosure sample makes the high disagreement among ESG rating agencies more important, as Christensen, Serafeim, and Sikochi (2021) find that high ESG disclosure leads to more disagreement.

ESG score disagreement

In a research paper focusing on ESG score divergence Lopez and Contreras (2020) discover that Thomson Reuters Eikon, RobecoSAM, and Sustainalytics only have highly consistent ESG scores for the worst-performing 10% of companies, with a strong correlation of 0.95 or higher. Furthermore, the correlation between ESG scores is less than 0.5 for the vast majority of companies. They also find that only two of the approximately ten categories that make up the E, S, and G pillars are equal among the three ESG rating agencies. Across sectors, the pairwise correlation varies greatly. The energy sector has the lowest correlation, while financials, technology, and cyclical consumer goods and services have the highest correlations. The sector dependent ESG scores are due to the varying emphasis on the environmental pillar across sectors, which the authors argue are the most difficult to measure. Berg, Kölbel, and Rigobon (2020) also investigates the divergence of ESG ratings. They argue that the low correlation across ESG rating agencies implies that funds using different ESG ratings would agree on stock picks to some extent, but not on all stocks. The authors contend that the ESG information reflected in the ESG scores can move prices; however, the low correlation between ESG scores from different agencies represents signal dilution. Consequently, the ability of ESG scores to move prices has been significantly diluted. Furthermore, although rating agencies claim that their ESG scores are industry-specific, the authors discovered that this was not the case to a large extent.

Both Lopez and Contreras (2020) and Berg, Kölbel, and Rigobon (2020) find the disagreement amongst different rating agencies to be substantial. The signal dilution pointed out by Berg, Kölbel, and Rigobon (2020) is interesting given that our sample in relation to Christensen, Serafeim, and Sikochi (2021) probably has an extra high disagreement. Investors state that they never rely on a single ESG score (SustainAbility, 2020). Even if some investors do, the consensus ESG score will almost certainly be what is reflected in prices in the aggregate in any case. We argue that the large disagreement across different ESG scores will make the use of single ESG scores inconsistent and make the results unnecessarily susceptible to random variations. Therefore, we argue that constructing a consensus ESG score based on the average of the seven different ESG scores in our sample will be our best option for finding that ESG scores are value relevant. However, the possibility of a high degree of variance in this consensus ESG score, as predicted by Christensen, Serafeim, and Sikochi (2021), is cause for concern. The problem with this is highlighted by Semenova and Hassel (2015). They use a US sample to study the convergent validity of the environmental performance metrics from the three rating agencies MSCI (former KLD), Thomson Reuters (former ASSET4), and Global Engagement Services (GAS). They find that the different rating agencies' ESG scores share dimensions, but they do not converge in aggregate. As a result, using an aggregated score may obscure the relationship between the variables examined in a study. In a newer study, Serafeim and Yoon (2021) find that a consensus ESG score made from MSCI, Sustainalytics, and Thomson Reuters can predict future news for the companies in their sample. They discover, however, that this ability disappears for a subset of companies in their sample where the three rating agencies that comprise their consensus ESG score disagree the most. Hence, it seems to be the case that high variance in the consensus ESG score diminishes the quality of a consensus measure.

To gain insight into how the significant disagreement across ESG rating agencies affects investment results, we will review previous studies that had this as part of their focus. In the context of the sharp increase in the use of ESG scores, Li and Polychronopoulos (2020) examine how the gap between ESG rating agencies affects investors. To demonstrate this, the authors used two popular rating agencies to create two distinct portfolios that are overweight in the top 50 percent ESG rated companies from each. This was done for the aggregated ESG score and for each of the three pillars. Even though the two portfolios appear to be based on similar criteria, their investment outcomes are only modestly correlated or unrelated. The effect is more pronounced for the individual pillars, and with more prominent differences in Europe than in the US. The authors conclude that this is caused by differences in the individual company ESG scores from the two different providers. In another study, Boffo and Patalano (2020), under the responsibility of the OECD Secretary-General, issued a comprehensive report on ESG investing. The authors formed several active portfolios based on ESG ratings from each of the five most well-known ESG rating agencies. In addition to the Sharpe ratio, they examined the portfolios' abnormal returns using the Fama French 5 factor model to control for factor biases. Only one of the five rating agencies' ESG scores produced a portfolio that generated a positive alpha from investing in higher-scoring ESG stocks. For the remaining four providers, the trend was reversed. The low-scoring ESG portfolios returned a positive alpha, while the high-scoring ESG portfolios did not. The secretariat studied the methodologies of the four ESG rating agencies Bloomberg, MSCI, Sustainalytics, and Thomson Reuters in order to understand why there are such large disparities in ESG ratings. Their investigation uncovered several individual choices made by the rating agencies that can explain why certain companies can receive wildly divergent ESG scores from different rating agencies. One factor highlighted was to which extent an ESG score reflected significant ESG issues. Gibson, Krueger, and Schmidt (2020) investigate the consequences of different beliefs among ESG rating providers and how this will affect future stock returns for S&P 500 firms. According to the empirical findings, the effect on stock returns is dependent on which of the E, S, and G pillars the rating agencies disagree on. If the disagreement pertains to the environmental pillar, it is regarded as a proxy for risk and yields a higher return. While disagreements on the social or governance pillar have a negative impact on stock returns. The findings also show that rating agencies disagree more for larger companies, firms without a credit rating, and less for the most profitable companies (Gibson, Krueger, & Schmidt, 2020). According to the findings of Boffo and Patalano (2020) and Li and Polychronopoulos (2020), using a single ESG score to form investment strategies or portfolios will result in returns that are highly dependent on the rating agency of choice.

Summarizing the key arguments for H1 – The value relevance of ESG

We anticipate that ESG scores will impact stock prices because they are used by investment managers who manage trillions of dollars in assets for portfolio creation, investment strategies, and trading. The majority of the literature we reviewed on the value relevance of ESG scores concluded that ESG scores are value relevant. However, because these studies were not conducted on a high-disclosure sample, we argue that a significant portion of the value relevance in their findings is most likely due to the reverse causality bias identified by Krüger (2015). As a result, we do not expect to find all the seven individual ESG scores in our sample to be value relevant.

Based on the high and unpredictable impact that ESG rating agency choice has on investment results, as demonstrated by Boffo and Patalano (2020) and Li and Polychronopoulos (2020), we argue that a consensus ESG score, calculated by taking the average of seven ESG scores in our sample will be the most consistent measure of value relevance. Furthermore, we argue that the consensus ESG score, will have the best capacity to impact prices in aggregate. The large disagreement among rating agencies, on the other hand, will act as a signal dilution on this capacity (Berg, Kölbel, & Rigobon, 2020). The dilution will be most likely be strongest on the part of the sample in which rating agencies disagree the most, consistent with Serafeim and Yoon's findings (2021). Thus, we expect to find the consensus ESG score to be less value relevant for this part of our sample. Lastly, we argue that the concerns raised by Semenova and Hassel (2015), that ESG scores from different rating agencies do not converge in aggregate are less of an issue in our study. This is due to the fact that our consensus ESG score is made up of ESG scores from six different rating agencies. As a result, we have formulated our first hypothesis as follows:

H1: The consensus ESG score is value-relevant, high disagreement between different ESG score providers reduces value relevance.

3.1.2 Investment strategies yielding superior performance

As discussed in section 2, there are two requirements for developing an investment strategy that generates superior returns by holding a long position in ESG stocks. To start, there must be a positive correlation between ESG performance and superior performance. Second, future superior ESG performance cannot be reflected in the price. However, as argued in section 2, the fact that ESG stocks appear to have outperformed for some time will probably increase the likelihood of long-term negative serial correlation as predicted by mean reversion (DeBondt & Thaler, 1985). That is, taking a short position in a specific category of these companies may prove to be more profitable than taking a long position. Thus, ESG stocks could be viewed as a group of stocks that have been raised by the tide. To expand on this analogy, the key question may be which companies will remain when the tide eventually recedes.

The link between ESG performance and financial returns will be the starting point for this sub-section. Friede, Busch, and Bassen (2015) wrote the research paper that provides the most comprehensive overview of this topic. They summarize the results of over 2000 empirical studies between 1970 and 2015 on the relationship between ESG criteria and corporate financial performance. According to the findings, approximately 90% of the studies find the correlation to be nonnegative. Furthermore, a large majority of studies observed a positive relationship, an effect which is more pronounced in North America. The observed relationship is also stable over time. The large number of studies, numerous methodologies, and datasets make the conclusion generalizable. In a much smaller meta-analysis, Wang, Dou, and Jia (2015) summarized the findings of 42 empirical studies on the relationship between corporate social responsibility (CSR) and financial performance. The authors' key finding is that there is a significant and positive association.

Based on the large number of studies included in Friede, Busch, and Bassen (2015), we will put the most weight on these results. According to the findings, it appears that ESG performance will have a non-negative effect on a company's returns in the vast majority of cases and will have a positive impact more often than not. As a result, we conclude that the positive relation is weak on average, while the non-negative link is strong.

According to S&P Global (2020), ESG investing is defined as using ESG information in investment analysis to invest sustainably while not sacrificing returns and, ideally, increasing profits. As we also saw in section 2, the two ESG screening strategies that have seen the most growth in recent years have been *general ESG integration* and *best-in-class*. Before discussing studies attempting to find the best strategies to achieve superior returns through ESG integration. It is crucial to highlight that the most severe challenge of attempting to achieve superior returns through ESG is that it necessitates a reduction in the investment universe, which we know from modern portfolio theory is not beneficial from a risk-return perspective. Pedersen, Fitzgibbons, and Pomorski (2020) demonstrate this by using the Markowitz portfolio optimization model to develop an ESG-efficient frontier as a theoretical framework for ESG investors. The authors use proxies for the individual E, S, and G pillars, and the MSCI ESG score for the combined ESG score. They discovered that removing 10% of the stocks with the lowest ESG scores reduces expected performance, resulting in a portfolio with a lower efficient frontier than a portfolio with no restrictions. By excluding stocks with the 20 percent lowest ESG scores, the efficient frontier was lowered even further. As a result, the authors concluded that the maximum Sharpe ratio achievable in a screened universe is lower than in an unscreened universe.

J.P. Morgan (2016) conducted a study to discover how ESG can help your portfolio. The study's key findings are that ESG can improve a portfolio's risk-return characteristics in three ways: lowering volatility, increasing Sharpe ratios, and limiting drawdowns. An important finding is that a high ESG score can cause exposure to traditional quantitative styles such as Size, Value, and Quality. However, after adjusting for these factors, it is concluded that ESG can still be a source of alpha in a portfolio context. Furthermore, the authors discover that investments based on the environmental pillar provide the highest returns in the United States, outperforming the aggregated ESG score, while investments based on the social and governance pillars provide the highest returns in Europe, but not enough to outperform the aggregated ESG score. Another interesting finding is that high and rising ESG factors outperform low and declining ESG factors. In another study Nagy, Kassam, and Lee (2015) investigate whether ESG can add alpha using MSCI ESG ratings. They use two methods to bring their theory to the test. Overweighting stocks with higher ESG ratings in strategy one, and in strategy two overweighting stocks that recently improved their ESG score. Both portfolios outperformed the MSCI World Index. Although some of the outperformance can be attributed to style factors such as momentum and small cap firms, factor biases can only explain a part of the outperformance. As a result, the authors conclude that investors that want to improve their ESG profile while taking some active risk can do so without compromising returns by employing this strategy. Giese et al. (2019-b) examine the methodologies for incorporating ESG into index-based portfolios, as well as how ESG incorporation can boost risk-adjusted returns in real-world index-based portfolios, using MSCI ESG scores. They construct two portfolios, one is a best-in-class strategy in which they select the top half of ESG-rated stocks in each region and sector, and the other is a strategy that favours companies with high ESG ratings and high ESG momentum. The best-in-class approach produced dramatically improved risk-adjusted returns due to higher price-to-book and price-to-earnings ratios, indicating higher valuations. After accounting for other factors, the best-in-class selection of companies with high ESG ratings was the primary source of active returns. Derwall et al. (2005) explore whether investing in US companies with high eco-efficiency scores from Innovest Strategic Value Advisors, which is essentially an ESG score with a focus on the environmental pillar, causes better portfolio results. According to their findings, companies with high eco-efficiency scores outperform stocks with low eco-efficiency scores. To demonstrate how the findings could be applied by investors in a more realistic manner, the authors used the best-in-class method to create portfolios that included companies with an eco-efficiency score above a pre-determined threshold. On a risk and style-adjusted basis, this portfolio outperformed a portfolio of companies with low eco-efficiency scores by 6%, constructed using the opposite of the best-in-class methodology, i.e., the worst-in-class methodology.

According to the findings of J.P. Morgan (2016), Nagy, Kassam, and Lee (2015), Giese et al. (2019-b), and Derwall et al. (2005), ESG investing can generate superior returns despite the obstacle of a restricted investment universe, as pointed out by Pedersen, Fitzgibbons, and Pomorski (2020). However, because each of these studies relies on a single ESG score, the results may be susceptible to random effects. This is because with 70 different ESG scores (Li & Polychronopoulos, 2020), creating one portfolio using each of these, some of these 70 portfolios will almost unavoidably generate superior performance simply by chance. However, this does not imply that these specific ESG scores will be able to consistently pick the stocks that will outperform in the coming years, which is what investors are always searching for (Grinold & Kahn, 2020). We argue that using the consensus ESG score from multiple ESG rating agencies will help diversify away random effects and lead to more consistent results.

Significant ESG issues

In recent years, evidence has emerged that companies that prioritize significant ESG issues outperform their peers, as briefly mentioned in section 2.2.2. Khan, Serafeim, and Yoon (2016) provide the most conclusive confirmation of this. They use KLD data to distinguish between investments in significant and non-significant sustainability issues. Applying the Sustainability Accounting Standards Board (SASB) definition of materiality to a dataset of more than 2,000 US companies over a 21-year period. The main finding is that companies that excel at significant sustainability issues earn higher returns than those that do not. Another intriguing finding is that achieving success on sustainability issues that are not defined to be significant, appears to have no impact on overall performance. The best-performing companies are those that perform poorly on sustainability issues that are not defined as significant while performing well on significant sustainability issues. Heijningen (2019) builds on the findings of Khan, Serafeim, and Yoon (2016) and investigates the impact of ESG on stock price returns, specifically how this impact is different when ESG issues are significant. The author builds active portfolios based on three different criteria. First, ESG scores on industry specific significant ESG issues. Second, not significant ESG issues and third, as a control, regular ESG scores not adjusted for whether ESG issues are significant or not. The classification of whether ESG issues are material or not, is done using the RobecoSAM materiality matrix. The results indicate that significant ESG issues strengthen the relationship between ESG performance and superior performance. In addition, their findings show that there is a range where significant ESG issues has a positive impact on abnormal returns; however, this is not a linear relationship. This is because there appears to be a threshold beyond which the portfolio ranked high on significant ESG issues is outperformed by the middle-ranked portfolio. In another study also building on the work of Khan, Serafeim, and Yoon (2016), Consolandi, Eccles, and Gabbi (2020) are perhaps able to provide an answer of why the relationship is not linear. In this study, the authors create indexes that allow them to determine how many and to what extent each significant ESG issue influences value drivers, as well as the concentration of these ESG issues within each industry. According to the findings, the market appears to reward companies operating in industries with a high concentration of significant ESG issues. However, the authors argue that overweighting an active portfolio against a high concentration of significant ESG issues serves as a risk concentration factor.

The findings of Khan, Serafeim, and Yoon (2016) can be illustrated with a simple example: Firms A and B operate in the same industry and are identical. They allocate the same quantity of resources to ESG issues. Firm A allocates resources mainly to ESG issues that have the potential to have a significant impact on enterprise value, i.e., significant ESG issues. Firm B, on the other hand, allocates resources primarily to ESG issues that do not have the potential to impact enterprise value, i.e., non-significant ESG issues. According to Khan, Serafeim, and Yoon (2016), this implies that ESG will have a positive and significant value relevance for firm A, but will not be value relevant for firm B. If we consider the findings of Khan, Serafeim, and Yoon (2016) in light of what was discussed in section 2.2.2, it makes sense that the value relevance of ESG is positive for firm A. Because firm A allocates resources only to ESG issues that effectively mitigate significant ESG risks and enable the firm to capitalize on ESG opportunities, which is the ESG performance that will most likely transmit through both the idiosyncratic and systematic channels, as demonstrated by the DCF model used by Giese et al. (2019-b), and explained in section 2.2.1. While the findings of Khan, Serafeim, and Yoon (2016) for firm A make sense, we argue that their findings for firm B cannot be correct. Their findings imply that for firms with high performance on non-significant ESG issues, ESG performance will have no effect on company performance, implying that ESG performance is not value relevant in this case. According to shareholder theory, this would be corresponding to a case of allocating resources to a purpose that does not benefit the shareholders (Friedman, 1970). We therefore argue that using a corporation's resources on ESG issues that is not significant ESG issues, should have a negative value relevance. ESG scores are investment products (Boffo & Patalano, 2020), and a high ESG score should, in a perfect scenario, reflect high performance on significant ESG issues. However, while rating agencies agree on this concept, there is little agreement on what constitutes a significant ESG issue within each industry. Therefore, we anticipate that if we compared the ESG scores our two fictional companies A and B would have received from different rating agencies, there would be no clear pattern indicating that one is better than the other based solely on the ESG scores. Although it is obvious that in a world without information asymmetry, an investor would prefer firm A over firm B. The motivation behind our research question is to use the value relevance of ESG scores to enable an investor to differentiate between firms A and B and use this insight to construct a portfolio overweight in companies in the same category as firm A, which should generate superior performance in theory. Using the same logic, companies in the same category as firm B could potentially generate superior performance with a short position.

Summarizing the key arguments for H2 – The best long position

The evidence for a strong and clear relationship between ESG performance and superior performance is relatively weak (Friede, Busch, & Bassen, 2015), and the reduction in the investment universe that comes with ESG investing makes it even more difficult to generate abnormal returns (Pedersen, Fitzgibbons, & Pomorski, 2020). Despite this, several studies have shown that ESG investment strategies can represent a source of alpha (Derwall et al., 2005; Nagy, Kassam, & Lee, 2015; J.P. Morgan, 2016; Giese et al., 2019). However, we argue that the results may be susceptible to random non-consistent factors because these research articles only used one ESG score and did not perform robustness checks using ESG scores from different rating agencies or alternatively a consensus ESG score.

According to the findings of Khan, Serafeim, and Yoon (2016), the link between ESG performance and superior performance could possibly shift from weak to strong if the ESG performance is on significant ESG issues. ESG issues are considered significant if they have the potential to have a significant impact on a company's value. If the consensus ESG score can explain changes in a company's value, it is determined to be value relevant. As a result, we argue that value relevance is a measure of the significance of ESG issues. Accordingly, we argue that firms with a high ESG score when ESG is value relevant have the best potential to generate superior returns. Given that our argument is correct, a long position would be the best investment strategy for companies that primarily allocate their resources to significant ESG issues, making the costs associated with achieving a higher ESG score justifiable because they represent a benefit to shareholders, ref. (Friedman , 1970). Because this is a group of companies that primarily allocates resources to ESG issues that have the potential to significantly impact enterprise value, this would represent firm A in our hypothetical example previously discussed. Therefore, the following hypothesis have been formulated:

H2: Long positions in stocks with a high consensus ESG score generate superior performance when ESG is value relevant.

Summarizing the key arguments for H3 – The best short position

Based on what we have covered in the thesis, the previous hypothesis was developed based on what we argue to be the best option for generating superior performance from a long position. The same procedure will be followed for the arguments leading up to this hypothesis, but with the goal of locating the best short candidate. Two main reasons will be presented to demonstrate why firms in the same category as firm B in our hypothetical example are the best shorting option.

According to Khan, Serafeim, and Yoon (2016), ESG performance is not value relevant for companies that have a high performance on ESG issues that do not have the potential to significantly impact enterprise value. We argue that allocating resources to ESG issues that do not benefit shareholders is synonymous with other unnecessary expenditures that destroy value, and thus must have a negative value relevance from the standpoint of shareholder theory. In contrast to the argument presented in H2, we argue that firms with a high ESG score when ESG is not value relevant have the least potential to generate superior returns. This is because the cost of achieving a higher ESG score is unjustifiable, and thus should reduce the firm's value (Friedman, 1970). As a result, an investor seeking superior returns should short these companies, which correspond to firm B in our previously discussed fictional example. However, there is a second reason why companies in the same category as firm B should be shorted, which we will discuss next.

Over the last two decades, both Europe and the United States have seen significant growth in sustainable investing, with growth accelerating after 2012 and reaching exponential proportions (Eurosif, 2018) (US SIF, 2020), contributing to a sharp increase in the use of ESG scores as an input in investment research. Because ESG investing strategies such as general ESG integration and best-in-class screening have seen the most growth in recent years, more institutional investors will have a clear tendency to overweight stocks with high ESG scores. As a result, we believe that there will be a high demand for companies with a high consensus ESG score, which will drive up the price. Serafeim (2019) demonstrates this mechanism by showing that the valuation premium charged for companies with a high MSCI ESG score increases as positive public sentiment momentum grows. Rating agencies are information intermediaries, and thus have the purpose of reducing information asymmetry. However, we argue that the substantial discrepancy between rating agencies on what constitutes a significant/material ESG issue implies that they to a point are incapable to fulfil their role as information intermediaries. As a consequence of information asymmetry, investors are unable to differentiate between companies that would represent firm A and firm B in our hypothetical example. As Berg, Kölbel, and Rigobon (2020) point out, this represents signal dilution, which reduces the ability of ESG scores to move prices, making it difficult for prices to reflect all information. Thus, the prices of companies in the same category as firm B are likely to rise significantly as well. Because stocks with a higher ESG score in general have been regarded as winners, they have probably had a positive medium to long-term serial correlation. As predicted by mean reversion, with a history of long-term positive serial correlation there is a greater likelihood that this will shift to a negative serial correlation (DeBondt & Thaler, 1985). Hence, we argue that companies of the same type as Company B have been overpriced as a result of their association with other companies of the same type as Company A that have demonstrated strong ESG performance on significant ESG issues. As a result, we argue that the fundamentals of companies in the same category as firm B will cause these companies to be the first to fall as a result of both overvaluation and mean reversion. Accordingly, the following hypothesis has been formulated:

H3: Short positions in stocks with a high consensus ESG score generate superior performance when ESG is not value relevant.

3.2 Test methodology

The objective of our test methodology is to answer our research question:

RQ: Could the value relevance of ESG scores be used to form investment strategies yielding superior performance?

The research question will be addressed through answering three hypothesises.

3.2.1 H1 – The value relevance of ESG

H1: The consensus ESG score is value-relevant, high disagreement between different ESG score providers reduces value relevance.

H1's primary goal is to confirm that ESG scores are value relevant, and its secondary goal is to determine how the substantial ESG score disagreement affects the value relevance of ESG scores. To achieve this, we employ the Collins, Maydew, and Weiss (1997) version of the Ohlson (1995) model, in which the consensus ESG score is used as the value relevant non-financial information, represented by Beta 3 in model 2 below.

$$Stock\ Price_{i,t} = \beta_0 + \beta_1 BVPS_{i,t,} + \beta_2 EPS_{i,t} + \beta_3 ESG_{i,t} + e_{i,t} \tag{2}$$

To address the impact of ESG score disagreement on value relevance, we expand the model with two more variables, as shown in model 3 below. Beta 4 is the dummy variable that takes the value 1 for the companies where the standard deviation between the seven different ESG scores are above the 75th percentile, and 0 otherwise. Beta 5 is an interaction variable that reflects how high disagreement among rating agencies affects the relationship between the consensus ESG score and the stock price. The models 4-6 is included to investigate whether the results are dependent on performance on the individual E, S and G pillars.

$$Stock\ Price_{i,t} = \beta_0 + \beta_1 BVPS_{i,t,} + \beta_2 EPS_{i,t} + \beta_3 ESG_{i,t} + \beta_4 DISAGREE_{i,t} + \beta_5 ESG_{i,t} \cdot DISAGREE_{i,t} + e_{i,t} \tag{3}$$

$$Stock\ Price_{i,t} = \beta_0 + \beta_1 BVPS_{i,t,} + \beta_2 EPS_{i,t} + \beta_3 ENV_{i,t} + \beta_4 DISAGREE_{i,t} + \beta_5 ENV_{i,t} \cdot DISAGREE_{i,t} + e_{i,t} \tag{4}$$

$$Stock\ Price_{i,t} = \beta_0 + \beta_1 BVPS_{i,t,} + \beta_2 EPS_{i,t} + \beta_3 SOC_{i,t} + \beta_4 DISAGREE_{i,t} + \beta_5 SOC_{i,t} \cdot DISAGREE_{i,t} + e_{i,t} \tag{5}$$

$$Stock\ Price_{i,t} = \beta_0 + \beta_1 BVPS_{i,t,} + \beta_2 EPS_{i,t} + \beta_3 GOV_{i,t} + \beta_4 DISAGREE_{i,t} + \beta_5 GOV_{i,t} \cdot DISAGREE_{i,t} + e_{i,t} \tag{6}$$

The consensus ESG score, represented by Beta 3, is considered value relevant if it has a significant positive or negative impact on stock price, implying that the coefficient is significantly different from zero at the 1%, 5%, or 10% level. We expect that to find that the coefficient Beta 5 is positive. Based on what was discussed above, we expect that ESG score disagreement will have a negative impact on the value relevance of the consensus ESG score. As a result, Beta 5 is expected to have a negative coefficient.

For the models 3-6 the null hypothesis will be rejected, if $\beta_3>0$ and $\beta_5<0$.

3.2.2 H2 – The best long position

H2: Long positions in stocks with a high consensus ESG score generate superior performance when ESG is value relevant.

We distinguish between two main types of investment strategies to generate superior returns, those that use a short position and those that use a long position. In H2, we will address the part of our research question that asks whether using the value relevance of ESG scores to take a long position will result in superior performance.

Because the goal is to find an investment strategy that generates superior performance, we use returns as the independent variable rather than stock price. We use annual returns instead of monthly returns because ESG scores are updated annually. As stated in the introduction to Section 3, the dependent variable in our research question employs the terminology: superior performance. We will use two levels to determine whether an investment strategy produces superior performance. The first threshold is met if the investment strategy significantly improves the risk-return relationship as measured by the Sharpe ratio. The second level, alpha, will be used as the ultimate measure of superior performance. We use this two-level superior performance classification mainly to enable us to capture the studied relationships on a more dynamic scale than a rigid alpha definition would allow.

ESG investing strategies frequently lead to exposure to traditional quantitative investment strategies, and factor biases are commonly controlled for in this context (Nagy, Kassam, & Lee, 2015; J.P. Morgan, 2016; Khan, Serafeim, & Yoon, 2016; Heijningen, 2019). Without controlling for factor biases, our conclusion would be weak because we would conclude that our investment strategy generates alpha when the alpha is attributable to quantitative strategies. Therefore, we will determine whether our investment strategy generates superior performance by calculating alpha using the Fama and French (2015) five-factor model, which we will extend with the liquidity factor developed by Pástor and Stambaugh (2003), as argued by Heijningen (2019). However, because the liquidity and momentum factors are unavailable on annual data, these two factors will be controlled for in section 5, where we use a different methodological approach that allows us to use monthly data.

The Ohlson model is a well-established model that has been shown to consistently explain a large portion of price fluctuations using only two sources of financial information: earnings and book value (Collins, Maydew, & Weiss, 1997). The model's residual will be smaller if the addition of the non-financial information represented by the consensus ESG score allows the Ohlson model to explain a larger portion of price fluctuations. We compute a variable equal to the difference between the Ohlson model residuals with and without the consensus ESG score, a variable we have labelled as the VR variable. When the VR variable is positive, it indicates that ESG reduces the model's residuals, implying that ESG is value relevant, i.e., analogous to what the literature defines as significant/material ESG issues. If the VR variable is negative, adding ESG to the Ohlson model increases the residuals, implying that ESG is not value relevant, i.e., analogous to what the literature defines as an unsignificant/immaterial ESG issue. ESG is a factor that has explained fluctuations in stock price for companies with a positive VR variable; the more positive the VR variable, the more ESG has explained. ESG has not been found to be a factor in explaining stock price fluctuations for companies with a negative VR variable. On the contrary, other factors, most notably earnings and book values, are explaining these fluctuations. The more negative the VR variable, the stronger this effect is.

Superior returns - Sharpe ratio

Model 7 is used to test whether holding in stocks with a high ESG score when ESG is value relevant can improve the risk-return relationship of an investment strategy. Beta 2 is a dummy variable that has a value of 1 for companies with a positive VR variable and a value of 0 otherwise. Beta 1 is the effect of having a high ESG score, regardless of whether or not ESG is value relevant. Hence, it represents the impact of overweighting in stocks with a high ESG score, not considering if ESG is value relevant. Given that ESG is value relevant, Beta 2 is defined as the average effect of ESG on the Sharpe ratio. The interpretation of Beta 3 is the effect of having a high ESG score when ESG is value relevant on the Sharpe ratio. Implying that the ESG issues on which these companies excel, as evidenced by their high ESG scores, are significant ESG issues, cf. (Khan, Serafeim, & Yoon, 2016), making the costs associated with achieving a high ESG score justifiable, cf. (Friedman, 1970). As indicated by H2, we expect that Beta 3 will have a positive coefficient. Models 8–10 are attempting to determine whether the consensus E, S, and G pillars influence the outcomes. The rationale is that environmental issues are not equally essential for any company, which is an argument that cannot be applied as easily to the governance pillar because good governance is more universally important.

$$Sharpe\ ratio_{i,t} = \beta_0 + \beta_1 ESG_{i,t,} + \beta_2 posVR_{i,t} + \beta_3 ESG_{i,t} \cdot posVR_{i,t} + e_{i,t} \quad (7)$$

$$Sharpe\ ratio_{i,t} = \beta_0 + \beta_1 ENV_{i,t,} + \beta_2 posVR_{i,t} + \beta_3 ENV_{i,t} \cdot posVR_{i,t} + e_{i,t} \quad (8)$$

$$Sharpe\ ratio_{i,t} = \beta_0 + \beta_1 SOC_{i,t,} + \beta_2 posVR_{i,t} + \beta_3 SOC_{i,t} \cdot posVR_{i,t} + e_{i,t} \quad (9)$$

$$Sharpe\ ratio_{i,t} = \beta_0 + \beta_1 GOV_{i,t,} + \beta_2 posVR_{i,t} + \beta_3 GOV_{i,t} \cdot posVR_{i,t} + e_{i,t} \quad (10)$$

For the models 7-10 the null hypothesis will be rejected, with respect to superior performance as indicated by a significantly improved Sharpe ratio, if $\beta_3 > 0$.

Superior returns - Alpha

Model 11 is used to determine whether holding in stocks with a high ESG score when ESG is value relevant will result in an alpha-generating investment strategy. Beta 2-5 are exposure corrections for traditional quantitative investment strategies. The VR variable is the same as what was used in the Sharpe ratio models 7-10. Beta 6 is defined as the average effect of ESG on annual returns, with no distinction between a high and a low ESG score. Thus, this variable represents an investment strategy that selects stocks solely based on a high ESG score, regardless of whether ESG is value relevant for these companies. If ESG is value relevant, Beta 7 represents the average effect on annual returns with no distinction between a high or a low ESG score, i.e., high or low ESG performance. When ESG is value relevant and the companies have a high ESG score, Beta 8 determines the effect of this interaction on annual returns. We anticipate that Beta 8 will be greater than zero, as predicted by our hypothesis. Models 12–14 are included to determine if the results are influenced by the individual E, S, and G pillars.

$$(AR-RF)_{i,t} = \alpha_0 + \beta_1 Mkt-rf_{i,t} + \beta_2 SMB_{i,t} + \beta_3 HML_{i,t} + \beta_4 RMW_{i,t} + \beta_5 CMA_{i,t} + \beta_6 ESG_{i,t,} + \beta_7 posVR_{i,t} + \beta_8\ ESG_{i,t} \cdot pos\ VR_{i,t} + e_{i,t} \quad (11)$$

$$(AR-RF)_{i,t} = \alpha_0 + \beta_1 Mkt-rf_{i,t} + \beta_2 SMB_{i,t} + \beta_3 HML_{i,t} + \beta_4 RMW_{i,t} + \beta_5 CMA_{i,t} + \beta_6 ENV_{i,t,} + \beta_7 posVR_{i,t} + \beta_8\ ENV_{i,t} \cdot pos\ VR_{i,t} + e_{i,t} \quad (12)$$

$$(AR-RF)_{i,t} = \alpha_0 + \beta_1 Mkt-rf_{i,t} + \beta_2 SMB_{i,t} + \beta_3 HML_{i,t} + \beta_4 RMW_{i,t} + \beta_5 CMA_{i,t} + \beta_6 SOC_{i,t,} + \beta_7 posVR_{i,t} + \beta_8\ SOC_{i,t} \cdot pos\ VR_{i,t} + e_{i,t} \quad (13)$$

$$(AR-RF)_{i,t} = \alpha_0 + \beta_1 Mkt-rf_{i,t} + \beta_2 SMB_{i,t} + \beta_3 HML_{i,t} + \beta_4 RMW_{i,t} + \beta_5 CMA_{i,t} + \beta_6 GOV_{i,t,} + \beta_7 posVR_{i,t} + \beta_8\, GOV_{i,t} \cdot pos\, VR_{i,t} + e_{i,t} \quad (14)$$

For the models 11-14 will the null hypothesis be rejected, with respect to superior performance as indicated by alpha, if $\alpha_0 > 0$.

3.2.3 H3 – The best short position

H3: Short positions in stocks with a high consensus ESG score generate superior performance when ESG is not value relevant.

We distinguish between two main types of investment strategies to generate superior returns, those that use a short position and those that use a long position. In H3, we will address the part of our research question that asks whether using the value relevance of ESG scores to take a short position will result in superior performance. However, we do not use a short position when we evaluate the Sharpe ratio. Hence, we expect to observe that the investments strategy has a significant and negative effect on the Sharpe ratio.

Superior returns - Sharpe ratio

Model 15 is used to determine whether holding in stocks with a high ESG score when ESG is not value relevant will worsen an investment strategy's risk-return relationship. Beta 2 is a dummy variable that takes the value 1 for companies with a negative VR variable and 0 otherwise. Beta 1 is the effect of having a high ESG score regardless of whether or not ESG is value relevant. Beta 2 is defined as the average effect of ESG on the Sharpe ratio when ESG is not value relevant. When ESG is not value relevant and the companies have a high ESG score, Beta 3 determines the effect of this interaction on Sharpe ratio. Since we do not use a short position for the Sharpe ratio, we expect to find that Beta 3 is less than zero. Again, models 16-18 are included to evaluate if the results are affected by the individual E, S, and G pillars.

$$Sharpe\ ratio_{i,t} = \beta_0 + \beta_1 ESG_{i,t,} + \beta_2 negVR_{i,t} + \beta_3 ESG_{i,t} \cdot negVR_{i,t} + e_{i,t} \quad (15)$$

$$Sharpe\ ratio_{i,t} = \beta_0 + \beta_1 ENV_{i,t,} + \beta_2 negVR_{i,t} + \beta_3 ENV_{i,t} \cdot negVR_{i,t} + e_{i,t} \quad (16)$$

$$Sharpe\ ratio_{i,t} = \beta_0 + \beta_1 SOC_{i,t,} + \beta_2 negVR_{i,t} + \beta_3 SOC_{i,t} \cdot negVR_{i,t} + e_{i,t} \quad (17)$$

$$Sharpe\ ratio_{i,t} = \beta_0 + \beta_1 GOV_{i,t,} + \beta_2 negVR_{i,t} + \beta_3 GOV_{i,t} \cdot negVR_{i,t} + e_{i,t} \quad (18)$$

For the models 15-18 the null hypothesis will be rejected, with respect to superior performance in this case indicated by a significant and negative impact on the Sharpe ratio, if $\beta_3<0$.

Superior returns - Alpha

Model 27 is used to determine whether taking a short position in stocks with a high ESG score when ESG is not value relevant will result in an alpha-generating investment strategy. The VR variable is the same as the one used in Sharpe ratio models 15-18. Beta 7 is the average effect of ESG on annual returns if ESG is not value relevant, i.e., with no differentiation between a high and a low ESG score. When ESG is not value relevant and the companies have a high ESG score, Beta 8 determines the effect of this interaction on annual returns. Because we argue that the costs of achieving a high ESG score when ESG is not value relevant are unjustifiable, we anticipate Beta 8 to be greater than zero due to the short position. Again, models 20-22 are included to see if the results are affected by the individual E, S, and G pillars.

$$(-AR-RF)_{i,t} = \alpha_0 + \beta_1 Mkt-rf_{i,t} + \beta_2 SMB_{i,t} + \beta_3 HML_{i,t} + \beta_4 RMW_{i,t} + \beta_5 CMA_{i,t} + \beta_6 ESG_{i,t,} + \beta_7 negVR_{i,t} + \beta_8\, ESG_{i,t} \cdot neg\, VR_{i,t} + e_{i,t} \tag{19}$$

$$(-AR-RF)_{i,t} = \alpha_0 + \beta_1 Mkt-rf_{i,t} + \beta_2 SMB_{i,t} + \beta_3 HML_{i,t} + \beta_4 RMW_{i,t} + \beta_5 CMA_{i,t} + \beta_6 ENV_{i,t,} + \beta_7 negVR_{i,t} + \beta_8\, ENV_{i,t} \cdot neg\, VR_{i,t} + e_{i,t} \tag{20}$$

$$(-AR-RF)_{i,t} = \alpha_0 + \beta_1 Mkt-rf_{i,t} + \beta_2 SMB_{i,t} + \beta_3 HML_{i,t} + \beta_4 RMW_{i,t} + \beta_5 CMA_{i,t} + \beta_6 SOC_{i,t,} + \beta_7 negVR_{i,t} + \beta_8\, SOC_{i,t} \cdot neg\, VR_{i,t} + e_{i,t} \tag{21}$$

$$(-AR-RF)_{i,t} = \alpha_0 + \beta_1 Mkt-rf_{i,t} + \beta_2 SMB_{i,t} + \beta_3 HML_{i,t} + \beta_4 RMW_{i,t} + \beta_5 CMA_{i,t} + \beta_6 GOV_{i,t,} + \beta_7 negVR_{i,t} + \beta_8\, GOV_{i,t} \cdot neg\, VR_{i,t} + e_{i,t} \tag{22}$$

For the models 19-22 will the null hypothesis be rejected, with respect to superior performance as indicated by alpha, if $\alpha_0 > 0$.

4. Data, variable construction, descriptive statistics, and simple correlations

4.1 Data

The data used in this thesis, as well as the processing steps taken to use the data in the following analysis, will be presented in this section. In addition, any assumptions made will be explained and reviewed.

The majority of the data comes from the suppliers Bloomberg and Thomson Reuters Eikon, with additional information added to supply the analysis, including the Fama-French factors obtained from the Kenneth French Data Library. The Fama French factors used are for developed markets, including the Liquidity factor as an additional sixth factor (French, 2021).

Cross-sectional time-series data from the Eikon Refinitiv Terminal and the Bloomberg Terminal were used. We removed data when cleaning the data for the reasons listed in Table 1. A total of 4,065 businesses were dropped from our sample.

Omitting data

The cross-sectional time-series data included 34,370 firm year observations. To ensure that the results obtained from the models are in fact correct, a procedure to omit observations were undergone. This process includes omitting variables with not enough data to produce a significant result, or data that would affects the results significantly. The process also omitted variables if a country or sector had too few observations in it to obtain any results from that sector.

Table 1 shows the procedure followed prior to the use of the data. As shown in the Table 1 below, the number of companies has been reduced from 4,910 to 845, yielding 5 915 firm-year observations.

Omitted variables

Number of firm year observations: 5 915	
Companies with ESG score (2013 - 2019)	Total
All companies	4910
Less: Missing economic variables for all 7 years	*544*
Companies with economic variables for at least one year	4366
Less: Missing country variables (<5 observations)	*10*
Companies with at least 5 observations per country	4356
Less: Companies without a specified industry	*62*
Companies with specified industries	4294
Less: Missing Refinitiv ESG and pillar score for all 7 years	*57*
Companies with Refinitiv ESG scores for at least one year	4237
Less: Missing Sustainalytics ESG and pillar score for all 7 years	*3020*
Companies with Sustainalytics scores for at least one year	1217
Less: Missing Bloomberg Disclosure score for all 7 years	*55*
Companies with Bloomberg Disclosure scores for at least one year	1162
Less: Missing RobecoSAM score for all 7 years	*70*
Companies with RobecoSAM scores for at least one year	1092
Less: Missing CDP Climate score for all 7 years	*238*
Companies with CDP scores for at least one year	854
Less: Missing ISS Quality score for all seven years	*8*
Companies with ISS scores for at least one year	846
Less: Removing outlier	*1*
Final number of companies used in dataset	845

Table 1:

Dataset screening history. This table provides a structure of how companies and variables were omitted from the dataset, with the reasoning behind every process. Economic variables include WAS, CMC, EPS and BVPS. All other economic variables did not have missing values and were therefore not omitted from the dataset. Outliers were checked by creating boxplots for each main variable. Only one firm year observation warranted removal, meaning that only one stock was removed as an outlier.

If a country in our data had fewer than five stocks, the country was removed from the dataset. The countries were removed from the dataset because having fewer than five stocks per country could impact or skew the regression analysis.

The sample countries are spread across 20 European and the US. Countries with more than 420 stocks, or approximately 10% of the dataset, only include the United States, which accounts for nearly 63 percent of our data. The United Kingdom is the second-largest country in our dataset.

As Francis et al (2004) do, missing values are noted as zero, and no value is deleted because a value is only missing in one time period. As a result, our dataset is a balanced panel dataset, with an equal number of observations in each time period for each stock.

If, on the other hand, a particular stock is missing data for a variable for the entire time period, the stock is removed from the dataset, see Table 1.

Outliers

A graphical method was used on each of the variables to ensure that there were no extreme values or outliers in the financial variables. The method included the use of scatterplots created to look at the distribution of the various variables. There is at least one outlier in a scatter plot if one point is further off the regression line than any other point. If a group of points is all at the same distance from the regression line, they are all outliers. There were no groups of outliers in our data, just a line of variables indicating a continuous range.

Over our seven-year period, one company had extreme values that were so negative that the decision was made to remove this company not to affect our analysis's reliability and accuracy. Nevertheless, we see that the variable EPS has many negative values; these are a large grouping of low values in 2018, and they are not flagged as outliers in the method and thereby not removed from the data. The main objective is that the data represents actual events and figures, as 2018 was overall a more negative year than the other years in our dataset for all financial variables in the dataset; this effect on our variables must be considered.

Complete dataset

The distribution of data between the US and Europe is comparable, with 47 percent of the data coming from US companies, and the remaining 53 percent is data from European countries.

We use the Global Industry Classification Standard (GICS) to verify that the data is diversified, and that the results are applicable for all sectors in both the US, and in Europe (S&P Global, 2018). This classification defines 11 different sectors. For these 11 sectors, the main part of the data, approximately 18 percent, is in the Industrials sector. We have the smallest percentage of data in the Energy sector, with approximately 4 percentage. Implying that the results of this model can more readily be generalized for the Industrials sector, and to not the same extent as the Energy sector.

As part of the analysis, the results from the value relevance regressions are used to create portfolios, which are created using split time periods to avoid forward-looking bias. To avoid unnecessary manipulation and errors in the main dataset, a new dataset in the form of panel data is created for each rolling regression period. For the robustness tests a subset of the data with the first four years is created to collect enough data to calculate the value relevance of ESG scores on monthly returns and Sharpe ratio.

All datapoints are in the dataset identified by ISIN number and year, to create a panel data identifier belonging to each firm year observation. This specialized ID is used to ensure the validity of the data, as well as when creating subsets of the data to use in other contexts other than the main models.

4.2 Variable construction

Consensus scores

The consensus scores are used on all the main models to estimate the effect of ESG scores on both stock prices and returns. First, an average score was calculated to create the consensus scores for both the ESG and individual pillars. This average is taken on each particular stock for each year as the equations below suggest with the "i,t" subscript. Next, the individual scores were retrieved from both the Bloomberg and Thomson Reuters Eikon platforms, where the Refinitiv ESG and Refinitiv Controversies scores were retrieved from the Thomson Reuters Eikon platform. Subsequently, the Sustainalytics, RobecoSAM, Bloomberg Disclosure, CDP Climate, Water, Forest score, and ISS Quality score were retrieved from the Bloomberg platform.

$$AVG.ESG_{i,t} = (REF.ESG_{i,t} + REF.ESG.C_{i,t} + SUS.ESG_{i,t} + ROB.ESG_{i,t} + DISC.ESG_{i,t} + CDP_{i,t} + ISS_{i,t})/7$$

$$AVG.E = (REF.ENV_{i,t} + SUS.ENV_{i,t} + ROB.ENV_{i,t} + DISC.ENV_{i,t} + CDP_{i,t})/5$$

$$AVG.S = (REF.SOC_{i,t} + SUS.SOC_{i,t} + ROB.SOC_{i,t} + DISC.SOC_{i,t})/4$$

$$AVG.G = (REF.GOV_{i,t} + SUS.GOV_{i,t} + ROB.ECON_{i,t} + DISC.GOV_{i,t} + ISS_{i,t})/5$$

High and Low ESG

Dummy variable accounting for whether the consensus ESG score is in the highest or lowest quartile of the entire dataset on consensus ESG scores. As there is 845 firm year observation, one quartile is approximately 211 of the observations.

$$HIGH.ESG = \begin{matrix} 1\ if\ consensus\ ESG\ is\ higher\ than\ the\ 75th\ percentile \\ 0 \end{matrix}$$

$$LOW.ESG = \begin{matrix} 1\ if\ consensus\ ESG\ is\ lower\ than\ the\ 25th\ percentile \\ 0 \end{matrix}$$

Disagreement

The disagreement variable, $DISAGREE_{i,t}$, is used in models three through six, and is calculated as thus. The standard deviation of the consensus ESG is calculated, and if the ESG score for a particular stock in a particular year is above the 75th percentile of the standard deviation, the variable will take on the value 1. This variable will in the models three through six function as a dummy variable, accounting for the disagreement in form of high standard deviation.

$$DISAGREE_{i,t} = \begin{matrix} 1\ if\ the\ SD\ of\ consensus\ ESG\ is\ higher\ than\ the\ 75th\ percentile\ of\ SD_{AVG.ESG} \\ 0 \end{matrix}$$

Trend ESG

For the robustness tests an estimated value of the consensus ESG is created. To create this value, we take the individual consensus scores for all 845 stocks in the years 2013, 2014, 2015, and 2016 and use the exponential smoothing algorithm of excel, to create a continuation of the ESG scores from the first four years of the dataset.

This continuation will be the $Trend\ ESG_{2017}$, and is used to sort the 845 stocks into the 211 with the highest and lowest ESG values according to this new variable, to obtain robustness results in a realistic setting.

VR variable

To estimate the value relevance of ESG scores on Stock price, we use the Ohlson model with and without the impact of ESG. The ESG score used in this regression model is the consensus ESG score, explained above. By subtracting the residuals from each other, the new variable VR is constructed. This variable will be used in both the main models and the robustness tests.

$$Ohlson: \quad Stockprice = \alpha + \beta_1 EPS + \beta_2 BVPS + e_{Ohlson}$$

$$With\ ESG: \quad Stockprice = \alpha + \beta_1 EPS + \beta_2 BVPS + \beta_3 AVG.ESG + e_{ESG}$$

$$VR = e_{ESG} - e_{Ohlson}$$

Positive and Negative VR

The VR variable were further categorized into a positive and negative dummy variable. These two dummy variables were used in the main tests as interaction variables, as well as to obtain realistic portfolios for the robustness tests.

Differentiating between positive and negative VR will enable the models to differentiate between stocks where the consensus ESG score is value relevant for the stock price or not.

$$Positive\ VR = POS.VR = \begin{matrix} 1\ if\ VR > 0 \\ 0 \end{matrix}$$

$$Negative\ VR = NEG.VR = \begin{matrix} 1\ if\ VR \leq 0 \\ 0 \end{matrix}$$

The instances where $VR = 0$, the Ohlson model with the ESG score included performed identically as the Ohlson model without the ESG correction, and there was no additional explanatory power created from the relationship between ESG and stock price, therefore these instances were sorted together with the negative VR values, as they both contain stocks where ESG does not provide value relevance for the stock price.

Trend VR

For the robustness tests an estimated value of the VR variable is created. To create this value, we take the individual VR scores for all 845 stocks in the years 2013, 2014, 2015, and 2016. These VR values are then put through the exponential smoothing algorithm of excel, to create a continuation of the VR values from the first four years of the dataset.

This continuation will be the $Trend\ VR_{2017}$, and is used to sort the 845 stocks into the 211 with the highest and lowest consensus ESG scores according to the interaction with this new variable, to obtain robustness results in a realistic setting.

As the Trend VR variable will function as a dummy variable, the actual numerical value of this variable does not affect the models or further calculations. This variable will be used to look for patterns in regard to a positive or negative link between ESG score and stock price, looking at the interactions between ether a positive or negative Trend VR together with the consensus ESG score.

Yearly Sharpe ratio

First the monthly Sharpe ratio is calculated per stock, for all the 845 stocks in the dataset. This is done by the following equation.

$$Sharpe\ ratio_{monthly} = \frac{Monthly\ returns - Monthly\ Rf}{Std.dev(Monthly\ returns)}$$

The monthly risk-free rate used to calculate the monthly Sharpe ratios is the risk-free rate from the Fama-French monthly dataset. This monthly Sharpe ratio was then annualized in the following way, to fit the yearly form of the dataset.

The monthly data is annualized using an annual percentage yield (APY) formula. This type of calculation is made, as both the yearly returns and yearly Sharpe ratios are used in a portfolio context, and this method of annualizing data best suits the nature of the method they are to be applied to.

$$\mathrm{APY} = (1 + \mathrm{r1}) * (1 + \mathrm{r2}) * (1 + \mathrm{r3}) * (1 + \mathrm{r4}) * (1 + \mathrm{r5}) * (1 + \mathrm{r6}) * (1 + \mathrm{r7}) * (1 + \mathrm{r8}) * (1 + \mathrm{r9}) * (1 + \mathrm{r10}) * (1 + \mathrm{r11}) * (1 + \mathrm{r12}) - 1$$

Using this formula to calculate the yearly Sharpe ratios, we get:

$$Sharpe_{YEAR\ a} = (Sharpe_{1.a} * Sharpe_{2.a} * Sharpe_{3.a} * \ldots * Sharpe_{12.a}) - 1$$

Where the 1 to 12 in the subscript refers to the different months, and a indicated the relevant year the returns are calculated for.

Yearly returns

In much the same way as the yearly Sharpe ratio is calculated, so is the yearly returns calculated. The monthly returns are obtained from Thomson Reuters Eikon service, where the identifiers created for each firm year observation serve as a link to any new data added to the dataset, including calculated yearly returns and Sharpe ratio. The annualising of the monthly returns, use the APY formula as shown above.

$$Returns_{YEAR\ a} = (return_{1.a} * return_{2.a} * return_{3.a} * \ldots * return_{12.a}) - 1$$

Where the 1 to 12 in the subscript refers to the different months, and a indicated the relevant year the returns are calculated for.

4.3 Descriptive statistics

The descriptive statistics of the cleaned dataset, including both the dependent and independent variables for the main models, is presented in Table 2 below. The table includes statistical measures such as mean, median, standard deviation, minimum and maximum values, as well as skewness and kurtosis in Panel A, and a correlation matrix of the main variables in Panel B.

Panel A of Table 2 shows that the mean consensus ESG score for our sample is 51.09, with a median of 52.53. Thus, the consensus ESG score has a very symmetrical distribution, with a close mean and median the actual values range from 0 to 87.78. This shows that the ESG score, which can rank from 0 to 100, is lower distributed, but since a consensus score from several suppliers is used that weights differently, this rank still represents a reasonable variation in the dataset for both companies with higher and lower scores. As can be seen from the table, the individual pillar scores have environmental, social, and governance, respectively average values of approximately 44, 44, and 48, which is somewhat lower than the main score. Mean median versions of the pillars close to the mean values are also the individual pillar scores in the data set symmetrically distributed. From the maximum and minimum values of the scores, the main score is somewhat lower than the maximum score of the pillar scores. The main score is lower than the pill scores at maximum and other values because there are more scores in this variable than those on individual pillars, as can be seen from the variable construction.

Descriptive statistics

Variable	N	Mean	Median	Std. Dev	Min	Max	Skewness	Kurtosis
	Panel A: Descriptive statistics							
Stock price	5915	544.65	311.67	1173.14	0.00	27129.37	10.79	160.05
EPS	5915	22.96	15.17	113.55	-6749.93	864.02	-38.25	2191.59
BVPS	5915	236.99	122.54	599.63	-229.96	14508.21	14.53	280.87
Consensus Environmental score	5915	44.42	45.50	21.36	0.00	93.07	-0.18	-0.70
Consensus Social score	5915	43.81	43.71	19.96	0.00	94.99	0.10	-0.69
Consensus Governance score	5915	47.94	48.41	16.07	0.00	90.98	-0.29	0.09
Consensus ESG score	5915	51.09	52.53	15.69	0.00	87.78	-0.56	0.56

	Panel B: Correlation matrix						
	1.	2.	3.	4.	5.	6.	7.
1. Stock price	1.0000						
2. EPS	0.8806*	1.0000					
	(0.0000)						
3. BVPS	0.8381*	0.8558*	1.0000				
	(0.0000)	(0.0000)					
4. Consensus Environmental score	0.8465*	0.7235*	0.7891*	1.0000			
	(0.0000)	(0.0000)	(0.0000)				
5. Consensus Social score	0.0445*	0.0370*	0.0293	0.0525*	1.0000		
	(0.0000)	(0.0000)	(0.0240)	(0.0000)			
6. Consensus Governance score	0.0477*	0.0588*	0.0377*	0.0353	-0.0406*	1.0000	
	(0.0000)	(0.0000)	(0.0000)	(0.0066)	(0.0000)		
7. Consensus ESG score	0.1054*	0.0870*	0.0720*	0.0912*	0.2790*	0.5544*	1.0000
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	

Table 2:

*This table shows the descriptive statistics (mean, median, standard deviation, minimum, maximum, skewness and kurtosis) in Panel A. Panel B give the correlation matrix of the main variables used in this study: Stock Price, Book value per share (BVPS), Earnings per share (EPS), Consensus Environmental, Social, Governance, and ESG scores. P-values in parentheses, * indicates significance at the 1 percent level.*

The financial variables used in the models do not have a symmetrical distribution in the data. For example, the share price has a mean of approximately 545 against a median of 311, which indicates that much of the data set is against the higher values. The difference indicates that the dataset consists mainly of companies that, on average, perform well in the stock market. However, since the range of the share price variable varies from zero to well over 27,000, this represents enough company variation to show reliable results when we look at the effect on stock prices. We see similar results for the variables Earnings per share and Book value per share, which are control variables in our model. These variables are somewhat more centrally distributed than the share price when looking at the mean and median differences. Further, these variables have a much more extensive range between maximum and minimum values.

From Table 2, we see that earnings per share have a very negative minimum value. However, as section 4.1 explains, all variables underwent outlier-testing, proving that the low minimum values of the financial variables, and their range, are not classified as an outlier. We do, however, see that the dataset for the financial variables is negatively skewed in 2018. Nevertheless, as there are no clusters of negative data, and they are not directly outliers, they are kept in the dataset to include variations in the data and models and not tamper too much with the data.

Skewness is a metric that determines how symmetric the distribution of data is around its mean value, signifying that the skewness of a normal distribution is zero (Brooks, 2014, p.66). The financial variables price, earnings per share (EPS), and book value per share (BVPS) are not equally skewed one way in our financial variables. The stock price and book value per share are positively skewed, indicating that the data is grouped towards the right and has a higher value. The Earnings per share variable is the most skewed variable in the dataset, with skewness of -38.25, which is not surprising with its low minimum value dragging the data towards a lower point. The ESG score and governance score are slightly positively skewed, and the environmental and social scores are slightly negatively skewed. These four variables have minimal skewness, as noted when describing the difference between the mean and variation in the variables.

Kurtosis is a measure of how fat the distribution's tails are. The kurtosis coefficient of a normal distribution will be negative 3 (Brooks, 2014, p. 66). Kurtosis is a measure of how heavy-tailed or light-tailed the data are in comparison to a normal distribution. Data sets having a high kurtosis are more likely to contain heavy tails or outliers. Light tails or a lack of outliers are typical in data sets with low kurtosis. From Table 2, we see that all ESG scores have a minimal kurtosis measure, implying a thinner tail than a normal distribution, which is typical for such symmetrically distributed data as a score between 0 and 100. The financial data all have much higher kurtosis measures, with EPS having a kurtosis of over 2,000. As explained in section 4.1, this financial variable is heavily affected by very negative values in 2018 but is still not considered an outlier, and therefore not omitted from the dataset. Higher kurtosis measures are considered common in economic time series, although the kurtosis for EPS is high.

A correlation matrix containing the correlation coefficients between the variables in our data set is presented in Panel B of Table 4. The relevant p-value for each coefficient is displayed in parenthesis, and the star beside the coefficients denotes a significance level at the one percent level. A correlation coefficient ranges from 1 to -1, with 1 indicating a perfect positive connection and -1 indicating a perfect negative correlation (Brooks, 2014, p.69).

Only one negative link exists between the variables, namely the social and governance score, suggesting that the social score is not as crucial in relation to governance performance. Except for the correlation coefficients between social score and BVPS; and the coefficient between governance score and environmental score, all the correlations are significant and positive. With correlation coefficients of 0.88 and 0.85, stock prices are highly correlated with EPS and BVPS, respectively, consistent with the value relevance approach, suggesting that these variables should have a strong correlation.

Multicollinearity

To investigate whether there are problems with multicollinearity in our dataset, and mainly in our independent variables, a Variance Inflation Factor (VIF) has been used. The VIF tests how much of the change in one variable is dependent on another. In other words, how much the variation of the left-side variable is affected or determined by the correlation between the explanatory variables (Woolridge, 2016, p.86).

The result from the VIF test gives quite expected results. There is multicollinearity between several variables in all our models, mainly because the correlation between dependent variables that are a product of each other creates structural multicollinearity. However, since this is part of our research design, and a part of the tests we perform is to look at how coherent variables affect, among other things, stock prices. For example, Consensus ESG and Disagreement have moderate multicollinearity, which is expected as disagreement is a product of the same ESG scores that settle Consensus ESG.

Furthermore, we also find moderate multicollinearity between Fama-French factors HML, CMA, and RMW in all our models, which use Fama-French factors. Presumably, this is multicollinearity derived from our dataset. If these variables were something to use further in our models, they would be taken care of, but we only use them as control variables. Thus, multicollinearity does not affect the model's predictive power, and we choose not to modify the dataset to comply with this.

Heteroskedasticity

The error term in a model is heteroskedastic when the error variance is not constant (Brooks, 2014, s. 182). Suppose the errors are heteroskedastic and not treated correctly. In that case, we will not have Best Linear Unbiased Estimators (BLUE), and the standard errors might be incorrect and affect the results (Brooks, 2014, s. 183). The Breusch-Pagan general test for heteroskedasticity is applied to determine if our data is heteroskedastic and adjustments are needed. In addition to these tests, displaying the estimated residuals against one explanatory variable in a scatterplot might indicate if the errors are heteroskedastic. There is no discernible pattern in this scatter plot, implying that the errors are homoscedastic. Therefore, the null hypothesis of homoskedasticity and constant variance should be rejected, according to the results of the Breusch-Pagan tests in R. These findings indicate that our data is heteroskedastic. We employ heteroskedasticity-consistent standard error estimates, also known as robust standard errors, to deal with the heteroskedasticity in our data. As a result, the regression becomes more conservative, and more evidence is required to reject the null hypothesis (Brooks, 2014, p. 186). In all our tables, we present robust standard errors to verify sufficient evidence to support our hypothesis.

Descriptive statistics of VR

Table 3 depicts shows the distribution of the variable VR by country and sector. We observe that dividing into a variable based on residual difference does not affect data distribution at either country or sector level. On the other hand, both the distribution of positive and negative VR is very balanced and evenly distributed across industries and countries. The advantage of balancing the division is that any effects observed from such a variable will be unaffected by country or sector-specific events. Any findings will be generalized across countries and sectors. As expected, the USA and the UK, as the largest contributors to data, also have the largest number of observations at each level. One then sees that the distribution well represents the number of companies in the primary data set distributed by the size of a country.

Country and sector affects on VR

Country	Panel A: Caterogized by country			
	n	Average VR	Positive VR	Negative VR
Austria	*70 (1.2%)*	14.43*	42 (1.4%)	28 (1.0%)
Belgium	*56 (0.9%)*	-22.48	24 (0.8%)	32 (1.1%)
Denmark	*126 (2.1%)*	22.70*	77 (2.6%)	49 (1.7%)
Finland	*91 (1.5%)*	-34.59	36 (1.2%)	55 (1.9%)
France	*525 (8.9%)*	-15.94	226 (7.5%)	299 (10.2%)
Germany	*413 (7.0%)*	15.04*	227 (7.6%)	186 (6.4%)
Greece	*21 (0.4%)*	27.66*	16 (0.5%)	5 (0.2%)
Ireland; Republic of	*28 (0.5%)*	23.27*	20 (0.7%)	8 (0.3%)
Italy	*168 (2.8%)*	-9.09	88 (2.9%)	80 (2.7%)
Netherlands	*196 (3.3%)*	14.04*	114 (3.8%)	82 (2.8%)
Norway	*91 (1.5%)*	-0.99	47 (1.6%)	44 (1.5%)
Portugal	*35 (0.6%)*	-27.08	12 (0.4%)	23 (0.8%)
Spain	*175 (3.0%)*	-32.67	71 (2.4%)	104 (3.6%)
Sweden	*210 (3.6%)*	-21.92	95 (3.2%)	115 (3.9%)
Switzerland	*259 (4.4%)*	15.04*	154 (5.1%)	105 (3.6%)
United Kingdom	*665 (11.2%)*	-6.15	351 (11.7%)	314 (10.8%)
United States of America	*2786 (47.1%)*	0.62	1397 (46.6%)	1389 (47.6%)
Sector	**Panel B: Caterogized by sector**			
	n	Average VR	Positive VR	Negative VR
Communication Services	322 (5.4%)	-8.85	154 (5.1%)	168 (5.8%)
Consumer Discretionary	644 (10.9%)	-5.91	310 (10.3%)	334 (11.4%)
Consumer Staples	476 (8.0%)	4.29*	259 (8.6%)	217 (7.4%)
Energy	245 (4.1%)	-7.51	124 (4.1%)	121 (4.1%)
Financials	903 (15.3%)	4.49*	467 (15.6%)	436 (14.9%)
Health Care	532 (9.0%)	7.48*	284 (9.5%)	248 (8.5%)
Industrials	1071 (18.1%)	-1.49	548 (18.3%)	523 (17.9%)
Information Technology	553 (9.3%)	9.06*	311 (10.4%)	242 (8.3%)
Materials	525 (8.9%)	-14.77	235 (7.8%)	290 (9.9%)
Real Estate	259 (4.4%)	-7.92	121 (4.0%)	138 (4.7%)
Utilities	385 (6.5%)	-13.62	184 (6.1%)	201 (6.9%)

Table 3:

This table shows the effect of country and sector on the variable VR. Panel A depicts the Average VR, as well how many of each country has a positive or negative VR variable. Panel B depicts the same factors but divided by sector. The percentages besides each number in each category indicated the percentage of the total of that category, and the * indicates that the average VR for that country or sector is above the median VR for the entire dataset.

4.4 Simple correlations

As disagreement across different ESG score providers is a factor in our model, the correlation between the different suppliers is presented in Table 4.

As we can see from Table 4, just over 70 percent of the ESG scores are significantly correlated to each other, which is expected as they are mainly based on approximately the same approach on the same topic. We find that among the significant correlations, the relationship between the Refinitiv ESG score and the Bloomberg Disclosure score is the largest, with a coefficient of 0.66, which indicates that both are based on either the same data or a similar calculation method. Furthermore, we find that from the significant correlations, 2 of them are negatively correlated; Refinitiv Controversies score against RobecoSAM and CDP, which indicates that Refinitiv Controversies uses a different focus when calculating their score than the other two.

This correlation matrix indicates that, although there are significant correlations, there is still some disagreement among the providers, which will prove fascinating to test in the hypothesis.

Correlation of ESG scores

Variable	1.	2.	3.	4.	5.	6.	7.
1. Refinitiv ESG score	1.0000						
2. Refinitiv ESG-Controversies score	0.065*	1.0000					
	(0.0000)						
3. Sustainalytics ESG Risk Ratings	0.4875*	0.0322	1.0000				
	(0.0000)	(0.0133)					
4. RobecoSAM Sustainability score	0.3252*	-0.0616*	0.4008*	1.0000			
	(0.0000)	(0.0000)	(0.0000)				
5. Bloomberg Disclosure score	0.6571*	-0.0277	0.4378*	0.3254*	1.0000		
	(0.0000)	(0.0334)	(0.0000)	(0.0000)			
6. CDP Climate, Water, Forest score	0.4774*	-0.0887*	0.3428*	0.1997*	0.4998*	1.0000	
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)		
7. ISS Quality Score	0.0296	0.0708*	0.0494*	-0.0045	0.0131	0.0245	1.0000
	(0.0227)	(0.0000)	(0.0000)	(0.7274)	(0.3130)	(0.0592)	

Table 4:
*This table shows the correlation matrix for the ESG scores used in the consensus ESG variable. The p-values are in parenthesis under each coefficient, and the * indicates significance at the 1 percent level. The largest and most significant correlation is between Refinitiv ESG score and Bloomberg Disclosure score. Most correlations are positive and significant, although Refinitiv ESG-Controversies score is negatively correlated to most other scores.*

5. Robustness tests

5.1 H2 and H3 under realistic conditions

The empirical strategy used in the main tests on H2, H3, and H4 has a few vulnerabilities that we will address using the methods presented in this section. The first shortcoming is that it is not ideal to correct for exposure to quantitative investment strategies using annual data from only seven years. We were also unable to control for the quantitative styles' liquidity and momentum due to the use of annual data. Third, our models in the main tests have a forward-looking bias, and an investment strategy with no proven predictive ability is of less value to an investor in a real-world situation. Another weakness is that the Sharpe ratios used in the main tests is based on the average Sharpe ratios of each firm, rather than on the average returns from all stocks in a portfolio. All these concerns will be addressed by using a methodology that could be applied by investors in a more realistic manner. This will be accomplished by selecting stocks for portfolios that correspond to the investment strategies outlined in sections 3.2.2 and 3.2.3 for the two hypothesises H2 and H3. A strategy combining the long position in H2 with the short position in H3 will also be explored.

5.1.1 Portfolio construction methodology

We have ESG data from 2013 to 2019, to remove forward looking bias we use the data from 2013 to 2016 to estimate our model that will determine which stocks to hold from 2017 to 2020. Hence, we simulate an investor on December 31, 2016, who knows nothing about the future and has a four-year investment horizon. The investor will hold the same stocks from 01.01.2017 to 31.12.2020.

As explained in section 3.2.2, negative values for the VR variable implies that ESG is not value relevant, implying that other factors than ESG explain stock price fluctuations. Positive values for the VR variable, on the other hand, implies that ESG is value relevant. Two main arguments are made in this thesis: 1. Allocating resources to achieve a high ESG score when ESG is not value relevant, i.e., negative values for the VR variable, represents costs that does not benefit shareholders and should destroy value, cf. (Friedman , 1970). 2. Allocating resources to achieve a high ESG score when ESG is value relevant, i.e., positive values for the VR variable, on the contrary is costs that benefit shareholders, and hence have the potential to increase value, extrapolating Barney (2020) to ESG. The investment strategies in H2 and H3 are, as explained in section 3.2.2 and 3.2.3, based on an interaction between a high consensus ESG score and an interaction between negative values for the VR variable in H3 and positive values for the VR variable in H2. As a result, it is important that the companies that falls in the *ESG is not value relevant category*, i.e., negative VR, and the companies that falls in the *ESG is value relevant category*, i.e., positive VR, are consistent with the arguments summarized above. The companies that receive a positive and negative VR values are evenly distributed across sectors and countries, as shown in Table 3 in section 4.3. As each company receives a value for the VR variable for each of the 4 years between 2013 and 2016, an option is to use the average of these four numbers to determine which companies falls in each category. However, we argue that this could cause companies that have invested in ESG issues that have become value relevant at the end of the four-year period, such that these costs are justified, to be incorrectly categorized by a negative VR value, because the average of the four VR values in such cases frequently turns out negative. To avoid this mechanism, we have chosen to use a forecast function rather than the average to extend the current trend for each company, details in section 4.2. The number outputted by the forecast function will determine the value of the VR variable, that will be used to create the negative VR and positive VR dummy variables. Because both the negative and positive VR variables are dummy variables, the only thing that matters is whether the values are negative or positive, not the numerical values. Hence, selecting the companies in the 75^{th} percentile of the interactions represents the companies with the highest consensus ESG scores that also meets the criteria in the respective interaction in the models. The dataset contains a total of 845 companies; if divided into quartiles, this equates to approximately 211 companies in each quartile. As a result, selecting the 75^{th} percentile of interactions will result in portfolios containing 211 companies. In determining which of the ESG scores from the years 2013-2016 to use in the interaction, we use a method that is based on using the same forecasting function. Using the forecasting function effectively creates a trend line, such that companies that have improved their ESG score will receive a higher ESG score. Because the methodology selects the 211 companies with the highest consensus ESG scores, companies with high and rising consensus ESG scores will be overweighted. This rationale behind the choice of this method is based on the findings of J.P. Morgan (2016) who found that portfolios based on high and rising ESG scores was one of the best performing investment strategies.

Evaluating the portfolio returns

The monthly returns from each portfolio will be used as the independent variable in model 31, represented by r, to determine how much of our portfolio's returns can be explained by traditional quantitative investment strategies. Beta 1 represents a market exposure correction, Beta 2 represents a size correction, Beta 3 represents a book-to-market correction, Beta 4 represents a quality correction, Beta 5 represents an investment style correction, Beta 6 represents a momentum correction, and Beta 7 represents a liquidity correction.

$$(r-rf)_t = \alpha_0 + \beta_1 Mkt - rf_t + \beta_2 SMB_t + \beta_3 HML_t + \beta_4 RMW_t + \beta_5 CMA_t + \beta_6 MOM_t + \beta_7 LIQ_t + e_t \tag{31}$$

In model 31, the most important coefficient is α_0, if $\alpha_0>0$, the robustness tests support rejecting the null hypothesis. If the main tests indicate an alpha and the robustness test supports this, it validates a rejection of the null hypothesis. If the results from the main tests and the robustness tests provide different results, we will reach a decision based on an overall assessment. The portfolios four-year investment horizon's monthly returns will also be used to calculate the Sharpe ratio, so that we have two levels of determining superior performance as in the main tests. In the main tests we measured whether our strategy had a significant impact on the Sharpe ratio. In the robustness tests, a superior Sharpe ratio will be determined relative to how the investment strategy would have been conducted if the investor did not have access to the VR variable. This is due to the fact that the interaction between the VR variable and the level of the consensus ESG score is the main innovation of this thesis. Therefore, this comparison will reveal the incremental effect of this innovation on the Sharpe ratio when compared to the methodology that would have been used in its absence.

5.1.2 Robustness test – H2

H2: Long positions in stocks with a high consensus ESG score generate superior performance when ESG is value relevant.

Portfolio – high ESG scrore when ESG is value relevant

In H2, the goal is to construct a portfolio of companies with a high ESG score when ESG is value relevant. This will be accomplished by utilizing the interaction in model 11 outlined below. That is, choose the 211 stocks with the highest numerical value of this interaction. Because the positive VR variable is a dummy variable with a value of 1 if VR is positive and

0 if VR is negative, the numerical value of the interaction is entirely determined by the consensus ESG score. Thus, will our portfolio of 211 stocks represent the stocks with the highest ESG score when ESG is value relevant. In the absence of the VR variable, an investor using the same strategy would most likely invest in the 211 stocks with the highest consensus ESG score. Thus, this portfolio will be compared to the Sharpe ratio generated by the portfolio in H2, which we argue is the better long position. To evaluate if the results are affected, the same procedure will be applied to the individual E, S, and G pillars.

$$(AR - RF)_{i,t} = \alpha_0 + \beta_1 Mkt - rf_{i,t} + \beta_2 SMB_{i,t} + \beta_3 HML_{i,t} + \beta_4 RMW_{i,t} + \beta_5 CMA_{i,t} + \beta_6 ESG_{i,t,} + \beta_7 posVR_{i,t} + \boldsymbol{\beta_8\ ESG_{i,t} \cdot pos\ VR_{i,t}} + e_{i,t} \tag{11}$$

5.1.3 Robustness test – H3

H3: Short positions in stocks with a high consensus ESG score generate superior performance when ESG is not value relevant.

Portfolio – high ESG scrore when ESG is not value relevant

In H3, the goal is to construct a portfolio of companies with a high ESG score when ESG is not value relevant. This will be accomplished by using the interaction in model 19 outlined below. That is, exactly like in H2 choose the 211 stocks with the highest numerical value of this interaction. The negative VR variable is a dummy variable with a value of 1 if VR is negative and 0 otherwise. Therefore, is the numerical value of this interaction also entirely determined by the consensus ESG score. An investor without access to the VR variable trying to find the best short position in ESG stocks would most likely invest in the 211 stocks with the lowest consensus ESG score. Making a low consensus ESG score portfolio the best comparison, regarding the Sharpe ratio, to the portfolio in H3. The same procedure will be followed with the individual E, S, and G pillars to see if the results are affected by this.

$$(-AR - RF)_{i,t} = \alpha_0 + \beta_1 Mkt - rf_{i,t} + \beta_2 SMB_{i,t} + \beta_3 HML_{i,t} + \beta_4 RMW_{i,t} + \beta_5 CMA_{i,t} + \beta_6 ESG_{i,t,} + \beta_7 negVR_{i,t} + \boldsymbol{\beta_8\ ESG_{i,t} \cdot neg\ VR_{i,t}} + e_{i,t} \tag{19}$$

5.1.4 Robustness test – Combining H2 and H3

The rationale behind this third robustness tests is twofold. First, this test will detect if the strategies, when combined, can produce superior performance but not on an individual basis. Inferring that the asserted relationships are correct but do not have a strong enough impact to serve as the foundation for individual strategies. Second, if the two strategies can produce superior results on their own, a portfolio that combines these two investment strategies into a single investment strategy will perform even better. We were unable to test the combination in the main tests due to the dummy variable trap. The methodological approach used in the robustness tests, on the other hand, allows for this. Furthermore, it is consistent with our research question to pursue for superior performance by utilizing the value relevance of ESG scores. Therefore, we will construct a portfolio that combines a long position in the H2 portfolio with a short position in the H3 portfolio. An investor without access to the VR variable would most likely construct a long-short ESG portfolio by taking a short position in the 211 stocks with the lowest consensus ESG score and a long position in the 211 stocks with the highest ESG score. Consequently, this portfolio will be used to compare the Sharpe ratio.

6. Results

In this section, the findings are presented. The findings will be presented in the same format as the methodology. Accordingly, we will first present the results of the main tests before concluding, hypothesis by hypothesis, on each of the two levels of superior performance. The results of the robustness tests will be evaluated against the conclusions reached based on the main tests, allowing us to determine whether the robustness tests results provide validation of these conclusions. The section concludes with a summary that answers the research question.

6.1 H1 – The value relevance of ESG

H1: The consensus ESG score is value-relevant, high disagreement between different ESG score providers reduces value relevance.

6.1.1 Main test

We expect to find that the consensus ESG score is value relevant based on the material presented in section 3.1.1 and the conclusions drawn from it, and that high disagreement between different rating agencies will have a negative impact on this value relevance. The regression results from model 3, as explained in section 3.2.1, is presented in Table 5 below. As expected, are the consensus ESG is score positively and significantly associated with the share price, implying that the consensus ESG score is value relevant. In addition, high disagreement across ESG scores has a negative impact on the value relevance of the consensus ESG score. As stated in section 3.2.1, models 4-6 is included to determine whether the results are dependent on the individual E, S and G pillars. We anticipate that the governance pillar will have the highest value relevance because it is more universally important. The results of models 4-6 is presented in Table 6 below. These results indicate the same relationships as model 3. The governance score, as expected, is the most value relevant of the three pillars. Figure 1 illustrates how the disagreement across rating agencies have developed over time, as illustrated have there been a drop in disagreement after 2016.

H1 - Testing for value relevance on consensus ESG score	
	Stock price (Model 3)
Constant	-49.423
	p = 0.330
Earnings per share	3.086
	p = 0.333
Book value per share	1.103***
	p = 0.000
Consensus ESG	5.341***
	p = 0.00001
Disagreement	235.925***
	p = 0.009
Consensus ESG * Disagreement	-5.997***
	p = 0.004
N	5,915
R^2	0.404
Adjusted R^2	0.403
Residual Std. Error	906.350 (df = 5909)
F Statistic	799.813*** (df = 5; 5909)
Notes:	***Significant at the 1 percent level.
	**Significant at the 5 percent level.
	*Significant at the 10 percent level.

Table 5:
Model 3, and its regression results, depicted in the table, considers consensus ESG and stock price. The consensus ESG score is positively and significantly related to share price, showing that ESG scores are value relevant. High levels of disagreement reduce the value relevance of the consensus ESG score. At a 1% level of significance, the null hypothesis is rejected.

H1 - Testing for value relevance on individual pillar consensus scores

	Stock price		
	(Model 4)	(Model 5)	(Model 6)
Constant	89.558*	85.485	14.907
	p = 0.085	p = 0.131	p = 0.785
Earnings per share	3.101	3.106	3.087
	p = 0.333	p = 0.333	p = 0.336
Book value per share	1.105***	1.106***	1.105***
	p = 0.000	p = 0.000	p = 0.000
Consensus Environmental score	3.006***		
	p = 0.0005		
Consensus Social score		3.105***	
		p = 0.001	
Consensus Governance score			4.271***
			p = 0.00000
Disagreement	124.028**	152.185**	92.019
	p = 0.019	p = 0.011	p = 0.172
Consensus Environmental score * Disagreement	-4.589***		
	p = 0.003		
Consensus Social score * Disagreement		-5.442***	
		p = 0.001	
Consensus Governance score * Disagreement			-3.064**
			p = 0.046
N	5,915	5,915	5,915
R^2	0.402	0.402	0.402
Adjusted R^2	0.401	0.401	0.401
Residual Std. Error (df = 5909)	907.814	907.834	907.649
F Statistic (df = 5; 5909)	793.425***	793.340***	794.142***

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

Table 6:
Models 4, 5, and 6 are displayed in that order in the table, providing the model's results at a pillar level of ESG. Governance is, as one might expect, the most value relevant factor.

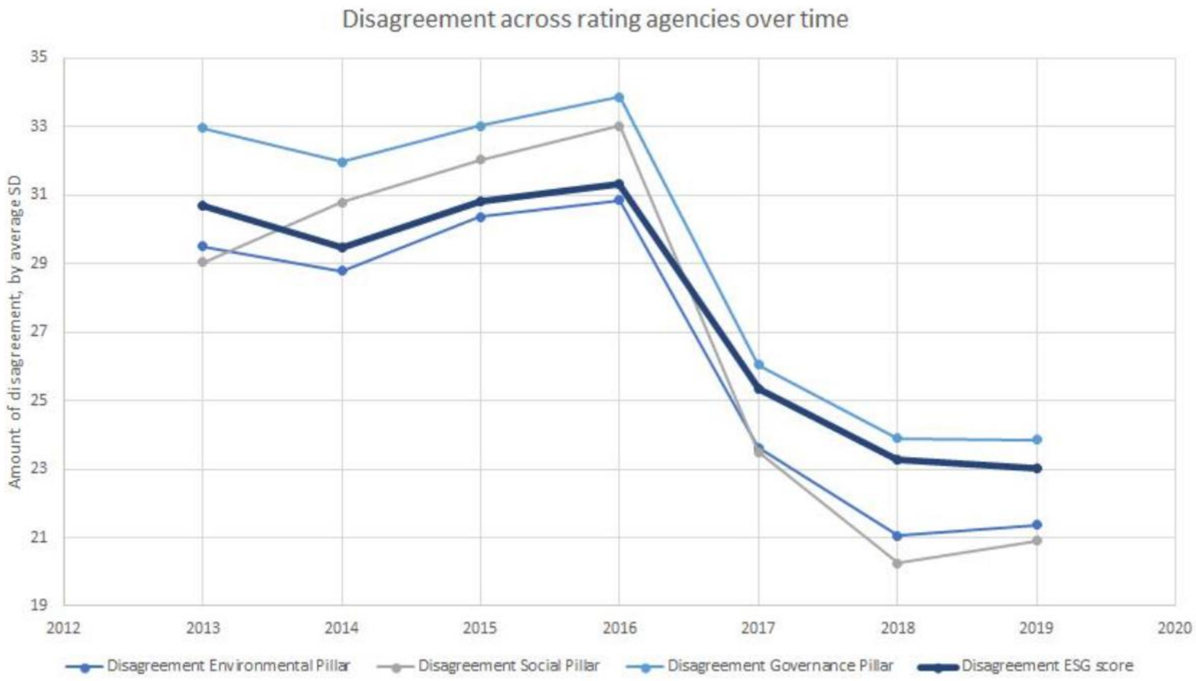

Figure 1:
The figure depicts the development of rating agency disagreement over time, particularly the decrease in disagreement since 2016. Disagreement is calculated by the standard deviation between the consensus scores and then aggregated to an average yearly disagreement level between the suppliers.

For the models 3-6 the null hypothesis will be rejected, if β_3>0 and β_5<0. Based on the results we find that β_3>0 and β_5<0 for the models 3-6 and reject the null hypothesis. The rejection is done at the 1 percent significance level for all the models 3-6 on both coefficients, with the exception of $\beta 5$ in model 6, which is rejected on a 5 percent significance level.

6.2 H2 – The best long position

H2: Long positions in stocks with a high consensus ESG score generate superior performance when ESG is value relevant.

6.2.1 Main tests

Superior returns – Sharpe ratio

Based on the material presented in section 3.1.2 and the conclusions drawn from it, we expect to find that a long position in the investment strategy represented by the interaction in model 7, explained in section 3.2.2, between consensus ESG score and positive VR to have a positive impact on the Sharpe ratio. The regression results from model 7 is presented in Table 7 below. As expected, is having a high consensus ESG score when ESG is value relevant positively and significantly associated with the Sharpe ratio, implying that the investment strategy improves the risk-return relationship. Models 8-10 are included once more to test whether the results are dependent on the individual pillars. The regression results from models 8-10 are presented in Table 8, these results indicate the same relationships as the primary model 7. However, we observe that a high consensus governance score has the greatest impact on the Sharpe ratio and explains more of the variation in Sharpe ratio, as indicated by the R squared. The fact that the consensus ESG score is negative in the absence of interaction with the VR variable implies that an investment strategy based solely on stocks with a high ESG score has a negative impact on the Sharpe ratio. Similarly, investing in stocks where ESG is value relevant, regardless of whether the companies have a high or low ESG score, will have a negative impact on the Sharpe ratio.

H2.1 - Investment strategy with a high ESG score when ESG is value relevant

	Yearly Sharpe ratio (Model 7)
Constant	5.946***
	p = 0.000
Consensus ESG score	-0.081***
	p = 0.000
Positive VR	-3.459***
	p = 0.000
Consensus ESG score * Positive VR	0.077***
	p = 0.000
R^2	0.040
Adjusted R^2	0.039
Notes:	***Significant at the 1 percent level.
	**Significant at the 5 percent level.
	*Significant at the 10 percent level.

Table 7:
The table presents the results for model 7 on how a positive value relevance affects the Sharpe ratio. In the absence of interaction with the VR variable, the consensus ESG score is negative, meaning that an investing strategy focused purely on stocks with a high ESG score negatively influences the Sharpe ratio. As the coefficient for the interaction term is significantly positive, we reject the null hypothesis at the 1% significance level.

H2.1 - Testing the investment strategy on individual pillar consensus scores

	Yearly Sharpe ratio		
	(Model 8)	(Model 9)	(Model 10)
Constant	3.869***	4.052***	5.424***
	p = 0.000	p = 0.000	p = 0.000
Consensus Environmental score	-0.048***		
	p = 0.000		
Consensus Social score		-0.053***	
		p = 0.000	
Consensus Governance score			-0.075***
			p = 0.000
Positive VR	-1.507***	-1.150***	-2.633***
	p = 0.00000	p = 0.0005	p = 0.000
Consensus Environmental score * Positive VR	0.047***		
	p = 0.000		
Consensus Social score * Positive VR		0.038***	
		p = 0.00000	
Consensus Governance score * Positive VR			0.064***
			p = 0.000
R^2	0.032	0.035	0.036
Adjusted R^2	0.032	0.034	0.036
Notes:	***Significant at the 1 percent level.		
	**Significant at the 5 percent level.		
	*Significant at the 10 percent level.		

Table 8:
The table presents the regression results of the models 8 to 10 and show that a high consensus governance score has the most significant impact on the Sharpe ratio and explains more of the variation in Sharpe ratio, as indicated by the R squared. In sum, we reject the null hypothesis based on these results.

For the models 7-10 the null hypothesis will be rejected, with respect to superior performance as indicated by a significantly improved Sharpe ratio, if β_3>0. Based on the results from the main tests we reject the null hypothesis at the 1 percent significance level for all the models 7-11. Inferring that an investment strategy with a long position in stocks with a high ESG score generates superior returns as measured by the Sharpe ratio when ESG is value relevant.

Superior returns – alpha

We expect an investment strategy based on the interaction in model 11, as explained in section 3.2.2, between the consensus ESG score and the positive VR variable to have a positive impact on annual returns based on the material presented in sections 2.2.2 and 3.1.2 and the conclusions drawn from it. On the one hand, based on the arguments presented in section 2.2.3 on markets being semi-efficient, we do not expect the investment strategy to generate alpha. However, on the other hand, could an innovative use of ESG scores that mitigates information asymmetry still be able to generate alpha. The regression results from model 11 is presented in Table 9 below. As expected, is having a high consensus ESG score when ESG is value relevant positively and significantly associated with annual returns. Furthermore, the investment strategy is able to generate alpha, with a p-value of 8.5 %. The models 12-14 is included to test whether the results are dependent on the individual pillars; the regression results are presented in Table 10. The interaction between individual consensus pillar scores and the positive VR variable remains positive, with the environmental pillar having the strongest link. Basing an investment strategy on the individual pillar scores are unable to generate alpha. However, as explained in the variable construction in section 4.2, this could be due to the fact that the consensus ESG score is based on 7 different ESG scores, while the pillar scores are based on 4-5 scores.

H2.2 - Investment strategy with a high ESG score when ESG is value relevant	
	Yearly Excess returns (Model 11)
Constant (alpha)	0.093*
	p = 0.085
MKT-RF	-0.007***
	p = 0.000
SMB	0.116***
	p = 0.000
HML	-0.011***
	p = 0.008
RMW	-0.091***
	p = 0.000
CMA	0.004
	p = 0.591
Consensus ESG	-0.004***
	p = 0.00000
Positive VR	-0.160**
	p = 0.011
Consensus ESG * VR	0.004***
	p = 0.001
R^2	0.449
Adjusted R^2	0.449
Notes:	***Significant at the 1 percent level.
	**Significant at the 5 percent level.
	*Significant at the 10 percent level.

Table 9:
The table presents the regression results for model 11, investigating the link between annual returns and positive value relevance. When ESG is value relevant, having a high consensus ESG score is positively and strongly associated with annual returns. With a p-value of 8.5 percent, the investing approach is also capable of generating alpha. We reject the null hypothesis; however, there is a risk of making a type 1 error due to the relatively large p-value.

H2.2 - Testing the investment strategy on individual pillar consensus scores

	Yearly Excess returns		
	(Model 12)	(Model 13)	(Model 14)
Constant (alpha)	-0.014	-0.001	0.040
	p = 0.721	p = 0.984	p = 0.437
MKT-RF	-0.007***	-0.007***	-0.007***
	p = 0.000	p = 0.000	p = 0.000
SMB	0.116***	0.116***	0.116***
	p = 0.000	p = 0.000	p = 0.000
HML	-0.011***	-0.011***	-0.011***
	p = 0.007	p = 0.008	p = 0.010
RMW	-0.092***	-0.092***	-0.092***
	p = 0.000	p = 0.000	p = 0.000
CMA	0.003	0.003	0.003
	p = 0.607	p = 0.622	p = 0.653
Consensus Environmental score	-0.003***		
	p = 0.00001		
Consensus Social score		-0.003***	
		p = 0.00001	
Consensus Governance score			-0.004***
			p = 0.00003
Positive VR	-0.092**	-0.036	-0.066
	p = 0.027	p = 0.412	p = 0.260
Consensus Environmental score * VR	0.003***		
	p = 0.0002		
Consensus Social score * VR		0.002*	
		p = 0.065	
Consensus Governance score * VR			0.002*
			p = 0.062
R^2	0.449	0.449	0.449
Adjusted R^2	0.448	0.449	0.448
Notes:	***Significant at the 1 percent level.		
	**Significant at the 5 percent level.		
	*Significant at the 10 percent level.		

Table 10:
Models 12, 13, and 14 are shown in the table, looking at the effect a positive value relevance has on the individual pillars of the ESG score. Individual consensus pillar scores positively interact with the positive VR variable, with the environmental pillar having the most substantial relation. Individual pillar scores alone are not enough to produce alpha in an investment strategy. Cannot reject the null hypothesis based on these results.

For the models 11-14 will the null hypothesis be rejected, with respect to superior performance as indicated by alpha, if $\alpha_0 > 0$. Based the results from the main tests we reject the null hypothesis for the primary model 11, because $\alpha_0 > 0$. This rejection is made with a p-value of 8.5 percent, implying that there is a risk of incorrectly rejecting the null hypothesis. We are unable to reject the null hypothesis for models 12-14.

6.3 H3 – The best short position

H3: Short positions in stocks with a high consensus ESG score generate superior performance when ESG is not value relevant.

6.3.1 Main tests

Superior returns – Sharpe ratio

Based on the material presented in section 3.1.2 we argue that an investment strategy based the interaction in model 15, presented in 3.2.3, between the consensus ESG score and the negative VR variable will have a negative impact on the Sharpe ratio.

H3.1 - Investment strategy with a high ESG score when ESG is not value relevant

	Yearly Sharpe ratio (Model 15)
Constant	2.487***
	p = 0.000
Consensus ESG	-0.004
	p = 0.548
Negative VR	3.459***
	p = 0.000
Consensus ESG * Negative VR	-0.077***
	p = 0.000
R^2	0.040
Adjusted R^2	0.039
Notes:	***Significant at the 1 percent level.
	**Significant at the 5 percent level.
	*Significant at the 10 percent level.

Table 11:
This table presents the regression results for model 15 that investigates the relationship between the Sharpe ratio and negative values for the VR variable. As expected, having a high consensus ESG score when ESG is not value relevant is negatively and significantly associated with the Sharpe ratio. Reject the null hypothesis at the 1% significance level.

The regression results from model 15, is presented in Table 11 below. As expected, having a high consensus ESG score when ESG is not value relevant is negatively and significantly associated with Sharpe ratio. Yet again, is models 16-18 included to test whether the results are dependent on the individual pillars. The regression results from models 16-18 are presented in Table 12, these results show the same relationships as the primary model 15. The governance pillar in the interaction with the negative VR variable has the strongest association with Sharpe ratio of the three pillars.

H3.1 - Testing the investment strategy on individual pillar consensus scores

	Yearly Sharpe ratio		
	(Model 16)	(Model 17)	(Model 18)
Constant	2.362***	2.902***	2.790***
	p = 0.000	p = 0.000	p = 0.000
Consensus Environmental score	-0.001		
	p = 0.807		
Consensus Social score		-0.015***	
		p = 0.003	
Consensus Governance score			-0.011*
			p = 0.060
Negative VR	1.507***	1.150***	2.633***
	p = 0.00000	p = 0.0005	p = 0.000
Consensus Environmental score * Negative VR	-0.047***		
	p = 0.000		
Consensus Social score * Negative VR		-0.038***	
		p = 0.00000	
Consensus Governance score * Negative VR			-0.064***
			p = 0.000
R^2	0.032	0.035	0.036
Adjusted R^2	0.032	0.034	0.036
Notes:	***Significant at the 1 percent level.		
	**Significant at the 5 percent level.		
	*Significant at the 10 percent level.		

Table 12:
The table presents the link between the Sharpe ratio and negative value relevance on the individual pillars of the consensus ESG score, depicting models 16, 17, and 18. The governance pillar in the interaction with the negative VR variable has the strongest association with the Sharpe ratio of the three pillars. Reject the null hypothesis at a 1% significance level.

For the models 15-18 the null hypothesis will be rejected, with respect to superior performance in this case indicated by a significantly lowered Sharpe ratio, if β_3<0. Based on the results from the main tests we reject the null hypothesis, because we find that β_3<0.

Superior returns – alpha

Based on the material presented in sections 2.2.2. and 3.1.2 we again argue that we expect a short position in an investment strategy represented by the interaction in model 19, explained in section 3.2.3, between the consensus ESG score and the negative VR variable to have a positive impact on the annual returns. As briefly discussed in H2, there is a strong case in the EMH for why we should not expect to find alpha, but there are also solid arguments for why we should find alpha through our innovative methodological approach. The regression results from model 19 is presented in Table 13. As expected, is having a high consensus ESG score when ESG is not value relevant negatively and significantly associated with annual returns, such that a short position makes the relationship positive. The investment strategy is close to generating a significant alpha with a p-value of 15.8 %. Therefore, will the robustness checks be more important to determine the ability of this investment strategy to generate alpha. Again, is models 20-22 included to test whether the results are dependent on the individual pillars. The regression results from models 20-22 are presented in Table 14. The investment strategy reflected by the interaction is positively and significantly associated with the returns for all the models 20-22 but is only significant at the 10 % level for the models using the social and governance consensus scores. However, as evident from Table 14 will the investment strategy be able to generate alpha using the consensus environmental score. Still, this may be due to a random effect because the consensus environmental pillar, as explained in section 4.2, is constructed from 5 different scores while the consensus ESG score is constructed from 7.

For the models 19-22 will the null hypothesis be rejected, with respect to superior performance as indicated by alpha, if α_0 >0. Based on the results from the main tests, we are can only to reject the null hypothesis for model 20. Although the primary model 19 is on the verge of producing a significant alpha, rejection would imply an unacceptably high risk of making a type 1 error.

H3.2 - Investment strategy with a high ESG score when ESG is not value relevant

	(-)Yearly Excess returns (Model 19)
Constant (alpha)	0.067
	p = 0.158
MKT-RF	0.007***
	p = 0.000
SMB	-0.116***
	p = 0.000
HML	0.011***
	p = 0.002
RMW	0.091***
	p = 0.000
CMA	-0.004
	p = 0.467
Consensus ESG	0.0005
	p = 0.568
Negative VR	-0.160**
	p = 0.017
Consensus ESG * Negative VR	0.004***
	p = 0.002
R^2	0.449
Adjusted R^2	0.449
Notes:	***Significant at the 1 percent level.
	**Significant at the 5 percent level.
	*Significant at the 10 percent level.

Table 13:
Model 19 investigate the relationship between annual returns and negative value relevance. The results show that having a high consensus ESG score when ESG is not value relevant is negatively and significantly associated with annual returns, and thereby a short position makes the relationship positive. The investment strategy is close to generating a significant alpha with a p-value of 15.8 %.

H3.2 - Testing the investment strategy on individual pillar consensus scores

	(-)Yearly Excess returns		
	(Model 20)	(Model 21)	(Model 22)
Constant (alpha)	0.106***	0.037	0.026
	p = 0.004	p = 0.323	p = 0.562
MKT-RF	0.007***	0.007***	0.007***
	p = 0.000	p = 0.000	p = 0.000
SMB	-0.116***	-0.116***	-0.116***
	p = 0.000	p = 0.000	p = 0.000
HML	0.011***	0.011***	0.011***
	p = 0.002	p = 0.002	p = 0.003
RMW	0.092***	0.092***	0.092***
	p = 0.000	p = 0.000	p = 0.000
CMA	-0.003	-0.003	-0.003
	p = 0.487	p = 0.505	p = 0.544
Consensus Environmental score	-0.0004		
	p = 0.499		
Consensus Social score		0.001*	
		p = 0.051	
Consensus Governance score			0.001*
			p = 0.071
Negative VR	-0.092**	-0.036	-0.066
	p = 0.033	p = 0.429	p = 0.283
Consensus Environmental score * Negative VR	0.003***		
	p = 0.0003		
Consensus Social score * Negative VR		0.002*	
		p = 0.072	
Consensus Governance score * Negative VR			0.002*
			p = 0.072
R^2	0.449	0.449	0.449
Adjusted R^2	0.448	0.449	0.448

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

Table 14:
This table shows models 20, 21, 22 and how negative value relevance in interaction with pillar consensus scores affects the annual returns. The models employing the social and governance consensus scores are only significant at the ten percent level. However, the investment strategy will create alpha utilizing the consensus environmental score, as evidenced by the findings. Can only reject the null hypothesis for model 20.

6.4 Robustness tests

In this section, we present the results of a different methodological approach used to test the robustness of the conclusions drawn from our main tests. The emphasis is on the two main investment strategies described in H2 and H3. Because the argued relationships for these strategies may not be enough to generate alpha as individual strategies, we will conduct a third robustness test that combines H2 and H3 into a single investment strategy.

6.4.1 H2 the best long position

Portfolio – high ESG scrore when ESG is value relevant

The investment strategy represented by the interaction in model 7, had a significant and positive impact on the Sharpe ratio at the 1% significance level. Accordingly, because the possibility of incorrectly rejecting the null hypothesis, type 1 error, in the main tests is small, we expect that the investment strategy will generate a superior Sharpe ratio under realistic conditions. On the contrary, the null hypothesis was rejected at a 10% significance level in the main test based on model 11. As a result, the likelihood of making a type 1 error increases.

A Sharpe ratio of 0.1187 is generated by an investment strategy that selects the 211 companies with the highest consensus ESG scores. While the investment strategy that invests in the 211 companies with the highest consensus ESG scores when ESG is also value relevant generates a Sharpe ratio of 0.2151, representing an increase of 80%.

As a result, we are less certain about the investment strategy's ability to generate alpha under realistic conditions. Table 15 displays the results of the H2 robustness test. Table 15 shows that an investment strategy based on selecting the 211 companies with the highest consensus ESG scores yields a Sharpe ratio of 0.1187. In comparison, the investment strategy that invests in the 211 companies with the highest consensus ESG scores when ESG is also value relevant generates a Sharpe ratio of 0.2151, representing an increase of 80%. In terms of alpha, the investment strategy in H2 is unable to generate alpha over the full period. However, at a 5% significance level, the strategy generates a significant and positive alpha in 2017, but also a significant and negative alpha in 2018. In 2018, 75% of the companies in our dataset had negative returns, compared to an average of 30% over the other 6 years. Because the proportion of companies reporting negative earnings per share did not change in 2018, it is possible that the negative alpha in 2018 was caused by a systematic exogenous risk factor.

Therefore, the investment strategy may have generated alpha over a different four-year time period excluding 2018. Table 16 presents the results of the robustness tests on the individual pillars. These findings are consistent with what we discovered in the main scores; an interesting discovery is that the environmental pillar generates the highest Sharpe ratio.

Robustness test H2 - Best long candidate under realistic conditions

	Avgerage monthly excess returns				
	Full period	2017	2018	2019	2020
Constant (alpha)	0.011	0.015**	-0.022**	0.007	0.025
	p = 0.193	p = 0.026	p = 0.049	p = 0.596	p = 0.593
MKT-RF	-0.112	0.285	1.125	-0.274	-1.425
	p = 0.683	p = 0.172	p = 0.112	p = 0.497	p = 0.387
SMB	0.010	-0.439	-2.473**	0.548	2.668
	p = 0.980	p = 0.178	p = 0.012	p = 0.342	p = 0.176
HML	0.452	0.151	-0.465	-0.970	0.853
	p = 0.367	p = 0.684	p = 0.411	p = 0.424	p = 0.572
RMW	-0.609	-0.255	-3.831**	1.721	-1.952
	p = 0.314	p = 0.421	p = 0.037	p = 0.331	p = 0.431
CMA	-0.195	-0.364	-0.230	-2.432	-0.756
	p = 0.811	p = 0.436	p = 0.869	p = 0.239	p = 0.755
MOM	0.023	0.034	-2.504**	-0.705*	0.784
	p = 0.939	p = 0.912	p = 0.044	p = 0.100	p = 0.468
LIQ	-0.014	0.145	0.489	0.005	0.830
	p = 0.969	p = 0.429	p = 0.232	p = 0.997	p = 0.737
R^2	0.070	0.599	0.900	0.747	0.683
Adjusted R^2	-0.093	-0.103	0.726	0.305	0.129
Sharpe ratio:					
Trend ESG * Positive VR	0.2151				
Trend ESG	0.1187				

Notes:

***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

Table 15:
This table shows the regression result from inputting the monthly returns of the portfolio, that is picked based on the highest consensus ESG scores when the VR variable indicates that ESG is value relevant, into the Fama French five-factor model extended with liquidity. This is done to test how much these factors explain the returns and to evaluate whether our investment strategy is able to generate alpha. Unfortunately, the investment strategy is unable to produce an alpha over the full period. However, a significant and positive alpha is generated in 2017.

Robustness test H2 - Best long candidate under realistic conditions

	Avgerage monthly excess returns				
	Full period	2017	2018	2019	2020
Constant (alpha)	0.011	0.015**	-0.022**	0.007	0.025
	p = 0.193	p = 0.026	p = 0.049	p = 0.596	p = 0.593
MKT-RF	-0.112	0.285	1.125	-0.274	-1.425
	p = 0.683	p = 0.172	p = 0.112	p = 0.497	p = 0.387
SMB	0.010	-0.439	-2.473**	0.548	2.668
	p = 0.980	p = 0.178	p = 0.012	p = 0.342	p = 0.176
HML	0.452	0.151	-0.465	-0.970	0.853
	p = 0.367	p = 0.684	p = 0.411	p = 0.424	p = 0.572
RMW	-0.609	-0.255	-3.831**	1.721	-1.952
	p = 0.314	p = 0.421	p = 0.037	p = 0.331	p = 0.431
CMA	-0.195	-0.364	-0.230	-2.432	-0.756
	p = 0.811	p = 0.436	p = 0.869	p = 0.239	p = 0.755
MOM	0.023	0.034	-2.504**	-0.705*	0.784
	p = 0.939	p = 0.912	p = 0.044	p = 0.100	p = 0.468
LIQ	-0.014	0.145	0.489	0.005	0.830
	p = 0.969	p = 0.429	p = 0.232	p = 0.997	p = 0.737
R^2	0.070	0.599	0.900	0.747	0.683
Adjusted R^2	-0.093	-0.103	0.726	0.305	0.129
Sharpe ratio:					
Trend ESG * Positive VR	0.2151				
Trend ESG	0.1187				
Notes:					***Significant at the 1 percent level.
					**Significant at the 5 percent level.
					*Significant at the 10 percent level.

Table 16:
This table shows if our best long candidate's individual pillar portfolio returns are explained by the Fama French five-factor model extended with liquidity. In addition, whether the investment strategy is able to generate alpha. The three portfolios are picked based on high pillar scores when the VR variable indicates the ESG is not value relevant. The results are consistent with the findings from the main tests, and the portfolios do not generate significant alpha.

We reach the following conclusions about superior performance on each of the two levels based on the results of the robustness tests and the main tests taken together:

1. ***Sharpe ratio.***

The Sharpe ratio was significantly and positively impacted by the investment strategy in H2, leading to a rejection of the null hypothesis in the main tests. Based on the results from the *robustness tests that demonstrated that the Sharpe ratio generated by the investment strategy in H2 was superior to the one produced by the incumbent methodology. This is because our innovative methodology of combining the VR variable with a high ESG score resulted in a Sharpe ratio improvement of 81%. Since the results of the robustness tests validate the conclusion drawn from the main tests. We reject the null hypothesis in H2.*

2. ***Alpha.***

In the main tests, we rejected the null hypothesis with an 8.5% chance of making a type 1 error. In the robustness test, the H2 investment strategy produced a positive alpha, but not at a significance level that allowed us to determine that alpha is significantly different from zero, as a rejection would result in a 19.3% chance of making a type 1 error. Overall, we contend that there is some evidence that the investment strategy described in H2 can generate alfa on its own, although the evidence is limited. We end up labelling the conclusion partially rejected based on an overall assessment.

6.4.2 H3 the best short position

The investment strategy represented by the interaction between the consensus ESG score and negative values for the VR variable, had a significant and negative impact on the Sharpe ratio at the 1 % confidence level. Accordingly, we based on the main test confidently rejected the null hypothesis in H2. Therefore, we anticipate that the Sharpe ratio, which is not calculated on the short version of H3, will be inferior to the best short candidate in the absence of the VR variable in the robustness tests. In this case, we argue that the obvious short position and best comparison would be a portfolio that shorts the 211 stocks with the lowest consensus ESG score. We used a short position to test whether the investment strategy generated alpha and were unable to reject the null hypothesis because this would imply a 15.8% chance of making a type 1 error. Consequently, we do not expect to find a positive and significant alpha in the robustness tests. However, in the main test, we found that model 20, which employs the consensus environmental pillar, can generate an alpha. Hence, there is evidence supporting that this portfolio will outperform the others also under realistic conditions. The results from the robustness tests, is presented in Table 17 for the consensus ESG scores and Table 18 for the individual pillar consensus scores. As shown in Table 17, the Sharpe ratio, calculated on a long position, of the investment strategy in H3 is, as expected, inferior to the short candidate we argued was the best comparison in the absence of the VR variable. We can also see from

Table 17 that the investment strategy in H3 is unable to generate a significant and positive alpha. The robustness test fails to replicate the alpha found in the main tests using the consensus environmental pillar, as shown in Table 18. Thus, the individual pillar robustness tests produce the same results as the overall score.

Robustness test H3 - Best short candidate under realistic conditions

	(-)Avgerage monthly excess returns				
	Full period	2017	2018	2019	2020
Constant (alpha)	-0.004	-0.005	0.019*	-0.003	-0.008
	p = 0.672	p = 0.486	p = 0.084	p = 0.877	p = 0.866
MKT-RF	0.078	-0.418	-0.428	0.241	1.842
	p = 0.803	p = 0.219	p = 0.493	p = 0.622	p = 0.316
SMB	-0.085	0.468	1.716**	-0.615	-3.623
	p = 0.843	p = 0.356	p = 0.041	p = 0.383	p = 0.111
HML	-0.689	-0.091	0.537	1.497	-1.625
	p = 0.226	p = 0.883	p = 0.360	p = 0.327	p = 0.344
RMW	0.440	0.602	2.852*	-3.170	1.197
	p = 0.518	p = 0.275	p = 0.088	p = 0.173	p = 0.650
CMA	-0.141	0.251	-0.330	0.946	1.310
	p = 0.878	p = 0.739	p = 0.818	p = 0.685	p = 0.624
MOM	-0.193	-0.322	2.076*	0.721	-1.290
	p = 0.567	p = 0.534	p = 0.078	p = 0.151	p = 0.294
LIQ	-0.081	-0.027	-0.421	-0.817	-0.676
	p = 0.843	p = 0.928	p = 0.303	p = 0.534	p = 0.802
R^2	0.087	0.633	0.863	0.688	0.740
Adjusted R^2	-0.073	-0.009	0.623	0.142	0.284
Sharpe ratio:					
Trend ESG * Negative VR	0.0788				
Low trend ESG	0.1320				

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

Table 17:
This table shows how the portfolio returns of our best short candidate are explained by the Fama French five-factor model extended with liquidity and whether it generates. The portfolio is picked based on the 211 companies with the highest consensus ESG scores the VR variable indicates the ESG is not value relevant. The Sharpe ratio of the investment strategy in H3, based on a long position, is, as expected, lower than the short candidate we stated was the closest comparison in the absence of the VR variable.

Robustness test H3 - Best short candidates under realistic conditions on pillars			
	(-)Avgerage monthly excess returns		
	Environmental	Social	Governance
Constant (alpha)	-0.004	-0.003	-0.004
	p = 0.697	p = 0.734	p = 0.678
MKT-RF	0.075	0.061	0.072
	p = 0.814	p = 0.851	p = 0.821
SMB	-0.088	-0.091	-0.042
	p = 0.841	p = 0.837	p = 0.925
HML	-0.762	-0.752	-0.730
	p = 0.190	p = 0.202	p = 0.209
RMW	0.411	0.437	0.450
	p = 0.554	p = 0.535	p = 0.517
CMA	-0.126	-0.183	-0.134
	p = 0.894	p = 0.848	p = 0.887
MOM	-0.222	-0.205	-0.229
	p = 0.520	p = 0.556	p = 0.506
LIQ	-0.110	-0.123	-0.052
	p = 0.792	p = 0.772	p = 0.900
R^2	0.094	0.096	0.085
Adjusted R^2	-0.065	-0.063	-0.076
Sharpe ratio:			
Trend E/S/G * Negative VR	0.0751	0.0671	0.0830
Low trend E/S/G	0.1246	0.1474	0.1428
Notes:			***Significant at the 1 percent level. **Significant at the 5 percent level. *Significant at the 10 percent level.

Table 18:
This table shows if our best short candidate's individual pillar portfolio returns are explained by the Fama French five-factor model extended with liquidity and if the portfolio generates alpha. The three portfolios are picked based on high pillar scores when the VR variable indicates the ESG is not value relevant. The table illustrates that the robustness test fails to replicate the alpha obtained in the main tests utilizing the consensus environmental pillar. The results of the individual pillar robustness tests are the same as the aggregate score.

We arrive at the following conclusions regarding superior performance on each of the two levels based on the results of the robustness tests and the main tests taken together:

1. ***Sharpe ratio.***

In the main test we found that the investment strategy in H3 had a significant and negative impact on the Sharpe ratio, and rejected the null hypothesis. Since the strategy in the *robustness test caused a 40% lower Sharpe ratio than the traditional short candidate, this validates the conclusion drawn from the main test. As a result, we reject the null hypothesis.*

2. Alpha.

In the main tests, we were unable to reject the null hypothesis using our model based on the consensus ESG score, but we were able to reject the null hypothesis using the consensus environmental pillar score. The robustness test, on the other hand, shows no evidence of the environmental pillar generating alpha on its own, as the p-value is roughly 70%. Hence, the robustness check provides strong evidence that contradicts the conclusion reached using the main test. Furthermore, it validates the failure to reject the null hypothesis conclusion reached for the consensus ESG score and the other two pillar scores. Overall, we conclude that we are unable to reject the null hypothesis in H2.

6.4.3 Combining H2 and H3

We now know that the relationships between the investment strategies in H2 and H3 were unable to generate significant and positive alpha under realistic conditions as separate strategies. However, there are strong indications that if the strategies are combined, they may be able to do so. To begin, the long candidate in H2 has a substantially higher Sharpe ratio, whereas the short candidate in H3 results in the lowest Sharpe ratio of all portfolios. Second, the strategy in H2 produces a significant alpha in the main tests and is close to producing a significant alpha in the robustness test. As a result, we argue that there is a good chance that combining the two strategies will result in an alpha-generating strategy, allowing us to indirectly validate the underlying relationships asserted in H2 and H3. The results of the third robustness test are shown in Table 19 below. We find that the investment strategy that combines H2 and H3 can generate a significant and positive alpha over the entire period, with a p-value of 1.7%. The Sharpe ratio generated by this portfolio is also the highest of all portfolios, at 0.339, whereas the portfolio based solely on H2 produced a Sharpe ratio of 0.215. The Sharpe ratios are negative for a long-short investment strategy based on an investment strategy that does not use the VR variable, and thus takes a long position in companies with a high ESG score and a short position in companies with a low ESG score. This implies that a high or low ESG score alone is a poor predictor of financial performance, as there is insufficient differentiation between high and low ESG scoring companies' financial performance. Furthermore, an R-squared of 0.092 for the full period indicates that the six-factor model is incapable of explaining the investment strategy's returns. However, R squared shifts to 0.9 in 2020, a year when companies were heavily influenced by the pandemic, indicating that the model can now explain a large proportion of the returns. In addition, all of the coefficients representing traditional quantitative investment strategies are significant in 2020, with the exception of the liquidity factor. However, we tested for multicollinearity using the variance inflation factor method, which demonstrates multicollinearity between the three factors HML, RMW, and CMA, implying that the investment strategy in practice may only lead to exposure to one of these factors in 2020, see section 4.3 for details. It is also worth noting that the investment strategy was able to eliminate the negative alpha seen in both the portfolios H2 and H3 in 2018. This could imply that our investment strategy has a better systematic risk profile, as the highest market beta is found in 2020 and is 1.6%. Table 20 presents the results for the individual E, S, and G pillar consensus scores, which all produce similar alphas as the portfolio constructed from the consensus ESG score.

Based on the results of this third robustness test, we have reached the following conclusion about the ability of the investment strategies in H2 and H3 to produce superior performance as measured by the Sharpe ratio and alpha:

1. ***Sharpe ratio.***

In terms of the Sharpe ratio, all of our tests confirm the proposed relationships. As a result, the null hypothesis is strongly and confidently rejected for H2 and H3.

2. ***Alpha.***

Overall, the main tests and the robustness test have failed to provide strong evidence of superior performance as measured by alpha when the strategies in H2 and H3 were used as individual strategies. This test, on the other hand, allowed us to indirectly validate the relationships on which H2 and H3 are based. This is because this robustness test was able to determine whether the combined impact of the two strategies generated a significant alpha.

Robustnesstest H2 and H3 - Best long-short candidate under realistic conditions

	Avgerage monthly excess returns				
	Full period	2017	2018	2019	2020
Constant (alpha)	0.007**	0.009	-0.004	0.005	0.016**
	p = 0.017	p = 0.389	p = 0.651	p = 0.311	p = 0.022
MKT-RF	-0.035	-0.132	0.697	-0.033	0.417*
	p = 0.707	p = 0.754	p = 0.253	p = 0.794	p = 0.054
SMB	-0.076	0.029	-0.757	-0.066	-0.956***
	p = 0.554	p = 0.965	p = 0.225	p = 0.706	p = 0.005
HML	-0.237	0.060	0.073	0.527	-0.772***
	p = 0.162	p = 0.943	p = 0.887	p = 0.207	p = 0.006
RMW	-0.169	0.346	-0.979	-1.450**	-0.755**
	p = 0.405	p = 0.625	p = 0.449	p = 0.044	p = 0.032
CMA	-0.336	-0.113	-0.560	-1.486*	0.554*
	p = 0.223	p = 0.913	p = 0.673	p = 0.059	p = 0.079
MOM	-0.170*	-0.289	-0.427	0.016	-0.506***
	p = 0.094	p = 0.681	p = 0.626	p = 0.886	p = 0.008
LIQ	-0.095	0.119	0.068	-0.812*	0.154
	p = 0.434	p = 0.770	p = 0.847	p = 0.061	p = 0.557
R^2	0.228	0.306	0.596	0.822	0.964
Adjusted R^2	0.092	-0.909	-0.112	0.510	0.900
Sharpe ratio:					
H2 portfolio - H3 portfolio	0.3390				
High - Low trend ESG	-0.1887				
Notes:					***Significant at the 1 percent level.
					**Significant at the 5 percent level.
					*Significant at the 10 percent level.

Table 19:
This table shows the portfolio returns when using a long-short portfolio based on consensus ESG scores and the Fama French five-factor model extended with liquidity. With a p-value of 1.7 percent, we discover that the investment strategy that combines H2 and H3 can provide a large and positive alpha over the entire period. Furthermore, the Sharpe ratio created by this portfolio is also the highest of all portfolios, at 0.339, whereas the Sharpe ratio generated by the portfolio based exclusively on H2 is 0.215.

Robustnesstest H2 and H3 - Best long-short candidates under realistic conditions on pillars			
	Avgerage monthly excess returns		
	Environmental	Social	Governance
Constant (alpha)	0.007**	0.006**	0.006**
	p = 0.023	p = 0.028	p = 0.034
MKT-RF	-0.037	-0.063	-0.061
	p = 0.690	p = 0.479	p = 0.493
SMB	-0.076	-0.115	-0.092
	p = 0.557	p = 0.351	p = 0.451
HML	-0.243	-0.259	-0.154
	p = 0.158	p = 0.115	p = 0.339
RMW	-0.200	-0.226	-0.219
	p = 0.331	p = 0.251	p = 0.260
CMA	-0.313	-0.324	-0.419
	p = 0.262	p = 0.224	p = 0.115
MOM	-0.171*	-0.165*	-0.177*
	p = 0.097	p = 0.094	p = 0.070
LIQ	-0.104	-0.138	-0.090
	p = 0.400	p = 0.243	p = 0.439
R^2	0.226	0.280	0.245
Adjusted R^2	0.091	0.154	0.113
Sharpe ratio:			
H2 portfolio - H3 portfolio	0.3194	0.2816	0.2769
High - Low average E/S/G	-0.1118	-0.1341	0.0315
High - Low trend E/S/G	-0.1701	-0.2820	-0.2615
Notes:			***Significant at the 1 percent level. **Significant at the 5 percent level. *Significant at the 10 percent level.

Table 20:
The table provides the results on a long-short portfolio, picked on individual consensus pillar scores, corrected for the Fama French five-factor model extended with liquidity. The alphas generated by portfolios constructed from individual pillar scores are nearly identical to those generated by the overall ESG score.

6.5 Summing up the results

RQ: Could the value relevance of ESG scores be used to form investment strategies yielding superior performance?

The process followed to answer the research question is presented in structured manner in Table 21 below. We initiate the process to answer the research question by determining whether ESG scores are value relevant. We find strong evidence that that the average of seven

ESG scores is value relevant and confidently rejects the null hypothesis. Using the value relevance of ESG scores as a primary input, we construct two investment strategies. Using a two-level approach, we determine whether the strategies generate superior performance. The Sharpe ratio is used as the first superior performance criterion, and alpha which is most difficult to achieve is used as the second criterion.

The investment strategy in H2 is intended to generate superior performance by taking a long position. Our findings are as follows: 1. We confidently reject the null hypothesis regarding the Sharpe ratio using both the main and robustness tests. 2. We reject the null hypothesis by a narrow margin in the main tests; we are close but unable to reject the null hypothesis in the robustness test. In terms of alpha, we decide to partly accept H2.

In contrast to H2, the investment strategy in H3 is intended to produce superior results by taking a short position. Our findings are as follows: 1. We find strong evidence to reject the null hypothesis regarding Sharpe ratio using both the main test and the robustness test. 2. The main test fails to reject the null hypothesis, and the robustness test validates this conclusion.

The goal of combining the two investment strategies in H2 and H3 is to run another robustness test on the argued relationships. The key rationale behind this is that even if the relationships we have argued for are correct, they may not be strong enough to generate superior performance as individual strategies. The following are our findings from the combination of the two: 1. The investment strategy generates a Sharpe ratio that is a 200% improvement over a portfolio constructed form selecting the 211 companies with the highest ESG score. Consequently, the null hypothesis is clearly rejected in terms of Sharpe ratio. 2. The strategy also produces a positive and significant alpha with a p-value of 1.7%, which represents our strongest proof of alpha in this thesis. As a result, the combination of the two strategies validates the existence of the relationships proposed in H2 and H3. Figure 2 depicts these relationships in a more intuitive manner, suggesting that a higher ESG score has a more positive impact on Sharpe ratio the more value relevant ESG is, as indicated by the lightest line. However, the darkest line in Figure 2 shows that when ESG is the least value relevant, the Sharpe ratio decreases in parallel with a higher ESG score.

Summarized results of our hypothesis

	Main test		Robustness test			
Hypothesis	Probability of Type1 error		Probability of Type1 error	Δ Sharpe ratio		Conclusion
H1: Value relevance of ESG	0.001%	Reject null hypothesis				Accept H1
H2.1: Sharpe ratio	0.000%	Reject null hypothesis		+ 81%		Accept H2.1
H2.2: alpha	8.431%	Reject null hypothesis	19.300%		Fail to reject null hypothesis	Partially accept H2
H3.1: Sharpe ratio	0.000%	Reject null hypothesis		- 40%		Accept H3.1
H3.2: alpha	15.800%	Fail to reject null hypothesis	67.200%		Fail to reject null hypothesis	Reject H3.2
Combining H2 and H3			1.700%		Given this combination, accept H2.2 and H3.2	RQ verification

Table 21:
This table is concluding and summarising the result of our hypotheses. Both from the main tests and the robustness tests. The probability of type 1 error is included to explain the trade-offs involved evaluating whether to rejecting or not rejecting the hypotheses. The type 1 error probability is obtained from the central variable of each hypothesis is tested for; for example, for H1 we show the p-value of the consensus ESG score coefficient, not the Disagreement variable as this is a secondary observation. Each of the hypotheses from H2-H3 is labelled in two stages to illustrate our two definitions of superior performance in a structured manner. The combination of H2 and H3 in the last row of the table proves that an alpha is created when the effects of both investment strategies are combined, allowing us to confirm our research question.

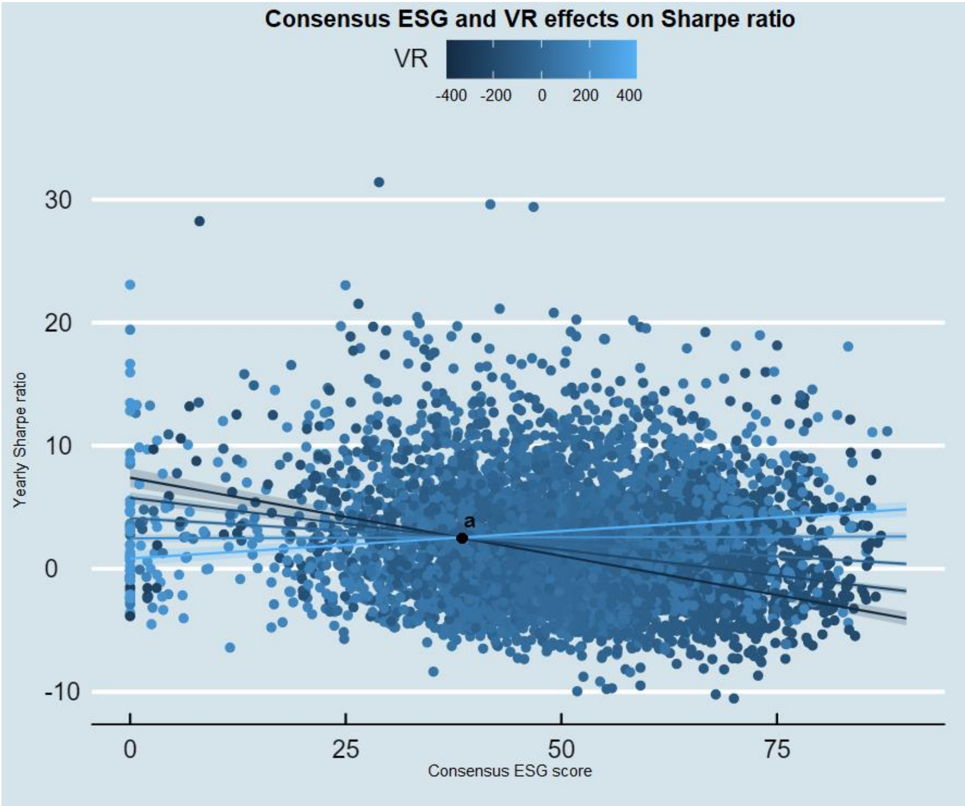

Figure 2:
This figure illustrates the relationship between the Sharpe ratio, the consensus ESG score, and value relevance (VR). The point market with an "a" indicates where the consensus ESG score goes from having a negative impact to a positive impact on the Sharpe ratio. The five lines indicates a regression line for when the VR variable takes on the values -400, -200, 0, 200 or 400, presenting the interaction between the VR variable and the consensus ESG score, where the colours correspond to the indicator at the top of the figure. When ESG is not value-relevant, the Sharpe ratio decreases with increasing ESG scores and vice versa when ESG is value relevant. When the ESG is not value relevant, ESG scores below 38.59, indicated by point "a", each unit increase of the ESG score is related to a -0.13 decrease of the Sharpe ratio. Implying that, for the lowest ESG scores, it is, of course, best that ESG is not value relevant. Accordingly, if a company have an ESG score below 38.59, and that the VR variable indicates that ESG is value relevant for this specific company, the figure clearly shows that allocating resources towards achieving a higher ESG score in this case would result in a higher Sharpe ratio. The opposite is true if the VR variable indicates that ESG is not value relevant for this specific company.

7. Conclusions

Our master's thesis is based on the simple idea that companies that allocate resources to matters with a minimal potential to affect the company's value are more likely to perform poorly than companies that allocate resources to matters with a very large potential to affect the company's value.

The objective of this paper is to investigate whether a methodological approach based on the value relevance of ESG scores can be used to develop investment strategies that generate superior performance. Based on the value relevance of ESG scores, we develop two investment strategies. First, as the core principle for the long position, we employ a methodology that allows us to determine which companies meet the following criteria: 1. A high consensus ESG score; 2. ESG is value relevant. Combining these two conditions, we claim, will result in a stronger relationship between ESG performance and corporate financial performance. The rationale for this claim is primarily based on the findings of Khan, Serafeim, and Yoon (2016) and Barney's (2020) argument that organizations should only focus on stakeholders who can have a direct impact on profit. Second, as the core principle for the short position, we employ the same methodology to determine which companies meet the following criteria: 1. A high consensus ESG score; 2. ESG is not value relevant. This strategy is based on shareholder theory, cf. Friedman (1970), and the counterargument of Barney (2020).

Our findings support the rationale and underlying factors that influenced the two strategies outlined above. This is because we find strong evidence that combining the two strategies into a single strategy generates superior performance. We use the following levels to determine whether the investment strategies yield superior performance: 1. the Sharpe ratio; 2. the Alpha. The latter is the Fama and French (2015) five-factor alpha extended with the liquidity factor (Pástor & Stambaugh, 2003) to control for potential exposure to factor biases. Our findings show no exposure to these factor biases over the full period. However, when looking at individual years, these factors go from explaining roughly 10% of returns in the full period to 90% in 2020, which we attribute to the global pandemic. In sum, we conclude that the value relevance of ESG scores can be used as a methodological approach to form investment strategies yielding superior performance.

ESG scores are investment products, and ESG rating agencies all state that their ESG scores reflect performance on ESG issues that have the potential to significantly impact enterprise value, most likely in an effort to make ESG scores more relevant for investors. However, if this were the case, a high ESG score would indicate improved corporate financial performance on its own. This thesis's findings clearly demonstrate that this is not the case. We argue that this is because there is disagreement about what constitutes a significant ESG issue. Previous research on this topic has typically relied on using some form of materiality map to get around this problem, as Khan, Serafeim, and Yoon (2016) and Heijningen (2019) have done. Because our methodology is much more efficient in identifying specific companies with high performance on ESG issues that have impacted their enterprise value. This is a significant contribution to the body of literature on this topic. Our findings, which support Khan, Serafeim, and Yoon's (2016) finding that high performance on significant ESG issues is positively associated with corporate financial performance, have added to the literature on the link between ESG performance and corporate financial performance by providing more evidence on this link. We also contribute to this literature by finding that high performance on non-significant ESG issues is negatively associated with corporate financial performance, which contradicts the findings of Khan, Serafeim, and Yoon (2016), who found a non-negative relationship. Because value relevance theory is central to our methodology, we also contribute to the literature, primarily by employing a consensus ESG score in conjunction with our dataset, which is less prone to reverse causality bias. The most practical implication of our thesis is that we provide ESG investors with an answer to the question: When is a high ESG score indicative of better financial performance?

One limitation of our thesis is that our methodology favours companies that emphasize ESG issues relevant to financial performance, and we label these companies peaches. However, it is possible that these companies will be labelled as lemons by stakeholders who have different priorities than investors looking to profit from ESG. Furthermore, we look at ESG issues that are value relevant within the timeframe of our dataset. Another limitation is that our dataset is dependent on data availability at our institution. Nevertheless, we noticed a clear trend in which our consensus ESG score became more value relevant as the number of ESG scores represented in the consensus increased. MSCI, Truevalue Labs, Truecost ESG Analysis, and S-Ray are examples of ESG scores that we wanted to include in our consensus ESG score but were unable to do so due to data constraints. Therefore, we believe that future research involving more of these new and technologically advanced ESG score providers has the potential to yield exceptional results.

8. References

Amel-Zadeh, A., & Serafeim, G. (2018). Why and How Investors Use ESG Information: Evidence. *Financial Analysts Journal 74 (3),* pp. 87-103.

Amir, E., & Lev, B. (1996). Value-relevance of nonfinancial information: The wireless communications industry. *Journal of Accounting and Economics 22 3 30*, pp. 3-30.

Aureli, S., Gigli, S., Medei, R., & Supino, E. (2019). The value relevance of environmental, social, and governance disclosure: Evidence from Dow Jones Sustainability World Index listed companies. *Wiley - Corporate Social Responsibility and Environmental Management*, pp. 1-10.

Barney, J. B. (2020). Measuring Firm Performance in a Way That is Consistent With Strategic Management Theory. *Academy of Management Discoveries Vol. 6, No. 1* , pp. 5-7.

Barth, M. E., & Clinch, G. (2009). Scale Effects in Capital Markets-Based Accounting Research. *Journal of Business Finance & Accounting 36 (3-4)*, pp. 253-288.

Beisland, L. A. (2009, April). A Review of the Value Relevance Literature. *The Open Business Journal* , pp. 7-27.

Berg, F., Kölbel, J. F., & Rigobon, R. (2020). *Aggregate Confusion: The Divergence of ESG Ratings*. MIT Sloan School Working Paper 5822-19.

Bloomberg . (2021). *Bloomberg Terminal NHH*.

Boffo, R., & Patalano, R. (2020). *ESG Investing: Practices, Progress and Challenges.* Paris: OECD.

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

Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of Finance, 52(1)* , pp. 57-80.

CDP. (2021). *CDP - Disclosure Insight Action* . Retrieved from Company Scores : https://www.cdp.net/en/companies/companies-scores

Christensen, D., Serafeim, G., & Sikochi, A. (2021). Why is Corporate Virtue in the Eye of The Beholder? The Case of ESG Ratings. *The Accounting Review https://doi.org/10.2308/TAR-2019-0506*.

Collins, D. W., Maydew, E. L., & Weiss, I. S. (1997). Changes in the value-relevance of earnings and book values over the past forty years . *Journal of Accounting and Economics 24*, pp. 39-67.

Consolandi , C., Eccles, R. G., & Gabbi, G. (2020). *How Material is a Material Issue? Stock Returns and the Financial Relevance and Financial Intensity of ESG Materiality.* Electronic copy available at: https://ssrn.com/abstract=3574547.

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. *Pac. Account. Rev 27.*, pp. 208-228.

De Villers, C., Naiker, V., & Van Staden, C. (2011). The Effect of Board Characteristics on Firm Environmental Performance . *Journal of Management Vol. 37 No. 6*, pp. 1636-1663.

DeBondt, W. F., & Thaler, R. (1985, July). Does the Stock Market Overreact? *The Journal of Finance Vol. 40 No. 3*, pp. 793-805.

Derwall, J., Guenster, N., Bauer, R., & Koedijk, K. (2005). The Eco-Efficiency Premium Puzzle. *Financial Analysis Journal Vol. 61 issue 2*, pp. 51-63.

Eccles, R. G., & Klimenko, S. (2019). *The Investor Revolution. Shareholders Are Getting Serious.* https://hbr.org/2019/05/the-investor-revolution: Harvard Business Review, May-June.

Eccles, R., & G., S. (2013). *The Performance Frontier: Innovating for a Sustainable Strategy.* Harvard Business Review, May 2013.

ESG and financial performance: aggregated evidence from more than 2000. (2015). *Journal of Sustainable Finance and Investment 5:4*, pp. 201-233.

Eurosif. (2016). *European SRI Study*. Brussel: European Sustainable Investment Forum.

Eurosif. (2018). *European SRI Study* . Brussel: European Sustainable Investment Forum.

Fama, E. F. (1970, May). Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance Vol. 25, No. 2*, pp. 383-417.

Fama, E. F. (1991, Dec). Efficient Capital Markets: ll. *The Journal of Finance Vol. 46 No. 5*, pp. 1575-1617.

Fama, E. F. (1998). Market efficiency, long-term returns, and behavioral finance. *Journal of Financial Economics 49*, pp. 283-306.

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

Fatemi, A., Glaum, M., & Kaiser, S. (2018). ESG performance and firm value: The moderating role of. *Global Finance Journal 38*, pp. 45-64.

Francis, J., LaFond, R., Olsson, P., & Schipper, K. (2004). Cost of equity and earnings attributes. *The Accounting Review, 79(4)*, 967-1010.

Freeman, E. R. (2010). *Strategic management: A stakeholder approach.* Cambridge University Press.

French, K. R. (2021, March). *Description of Fama/French 5 Factors for Developed Markets*. Retrieved from https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/f-f_5developed.html

Friede, G., Busch, T., & Bassen, A. (2015). ESG and financial performance: aggregated evidence from more than 2000 empirical studies. *Journal of Sustainable Finance & Investment, Vol. 5, No. 4*, pp. 210-233.

Friedman , M. (1970). The Social Responsibility of Business is to Increase its Profits. *The New York Times Magazine, September 13* .

Gibson , R., Krueger, P., & Schmidt, P. S. (2020). *ESG Rating Disagreement and Stock Returns* . European Corporate Governance Institute - ECGI Working Paper Series in Finance.

Giese, G., Lee, L.-E., Melas, D., Nagy, Z., & Nishikawa, L. (2019-a). Foundations of ESG Investing: How ESG Affects Equity Valuation, Risk, and Performance. *The Journal of Portfolio Management - volume 45 number 5*.

Giese, G., Lee, L.-E., Melas, D., Nagy, Z., & Nishikawa, L. (2019-b). Performance and Risk Analysis of Index-Based ESG Portfolios. *The Journal of Inedex-investing - ETFs, ETPs & Indexing Vol. 9 Num. 4*.

Graham, B., & Dodd, D. L. (2009). Discrepancies Between Price and Value. In *Security Analysis - Principles and Technique* (pp. 669-687). McGrawHill.

GRI & RobecoSAM. (2016). *Defining What Matters - Do companies and investors agree on what is material?* The Global Reporting in collaboration with RobecoSAM.

Grinold, R. C., & Kahn, R. N. (2020). *Advances in Active Portfolio Management - New Developments in Quantitative Investing.* McGrawHill.

Hassel, L., Nilsson, H., & Nyquist, S. (2005). The Value Relevance of Environmental Performance. *European Accounting Review, Vol. 14, No. 1*, pp. 41–61.

Healy, P. M., & Palepu, K. G. (2001). Information asymmetry, corporate disclosure, and the capital markets: A review of the empirical disclosure literature. *Journal of Accounting and Economics 31*, pp. 405-440.

Heijningen, K. v. (2019). *The impact of ESG factor materiality on stock performance of firms.* Erasmus Platform for Sustainable Value Creation.

Henriksson, R., Livnat, J., Pfeifer, P., & Strumpp, M. (2018). *Integrating ESG in Portfolio Construction* . Quantitative Management Associates LLC Newark NJ 07102: Stern School of Business Administration, New York University, and QMA.

ISS. (2021). *ISS*. Retrieved from ISS Analytics : https://www.issgovernance.com/solutions/iss-analytics/qualityscore

J.P. Morgan. (2016). *ESG - Environmental, Social & Governance Investing - A Quantitative Perspective of how ESG can Enhance your Portfolio.* J.P. Morgan - Global Quantitative & Derivatives strategy .

Jain, A. K., & Rezaee, Z. (2016). Value-Relevance of Corporate Social Responsibility:. *JOURNAL OF MANAGEMENT ACCOUNTING RESEARCH Vol. 28, No. 2*, pp. 29-52.

Jegadeesh, N., & Titman, S. (2001, April). Profitability of Momentum Strategies: An Evaluation of Alternative Explanations. *The Journal of Finance Vol. No. 2*, pp. 699-720.

Kaspereit, T., & Lopatta, K. (2016). The value relevance of SAM's corporate sustainability ranking and GRI sustainability reporting in the European stock markets. *Business Ethics: A European Review Vol. 25 Num. 1*, pp. 1-24.

Khan, M., Serafeim, G., & Yoon, A. (2016). Corporate sustainability: First evidence on materiality. *Accounting Review 91 (6)*, pp. 1697–1724.

Krüger, P. (2015). Corporategoodnessandshareholderwealth. *Journal of Financial Economics 115*, pp. 304-329.

Landau, A., Rochell, J., Klein, C., & Zwergel, B. (2020). Integrated reporting of environmental, social, and governance and financial data: Does the market value integrated reports? *Wiley - Business Strategy and Envirionment 29*, pp. 1750-1763.

Li, F., & Polychronopoulos, A. (2020). What Difference an ESG Rating Provider Makes! *Research Affiliates*.

Lo, A. W. (2007). *Efficient Market Hypothesis* . https://www.researchgate.net/publication/228319211.

Lopez, C., & Contreras, O. F. (2020, October). ESG Ratings: The Road Ahead . *SSRN Electronic Journal* , pp. 1-53.

Malkiel, B. G. (2019). *A Random Walk Down Wall Street - Completly Revised and Updated.* New York - London: W. W. NORTON & COMPANY.

Miralles-Quirós, M., Miralles-Quirós, J. L., & Valente Gonçalves, L. M. (2018). The Value Relevance of Environmental, Social, and Governance Performance: The Brazilian Case. *Sustainability*, pp. 1-15.

Morningstar. (2021, January 28). *A Broken Record: Flows for U.S. Sustainable Funds Again Reach New Hights*. Retrieved from Morningstar: https://www.morningstar.com/articles/1019195/a-broken-record-flows-for-us-sustainable-funds-again-reach-new-heights

MSCI. (2020). *MSCI ESG Ratings Methodology* . MSCI ESG Research .

Nagy, Z., Kassam, A., & Lee, L.-E. (2015). Can ESG add Alpha? An Analysis of ESG Tilt and Momentum Strategies . *MSCI ESG Research Inc.*

OECD. (2017). *Investment governance and the integration of environmental, social and governance factors* . OECD.

Ohlson, J. A. (1995). Earnings, book value, and dividends in equity valuation. *Contemporary Accounting Research, 11 (2)*, pp. 661-687.

Pástor, Ľ., & Stambaugh, R. F. (2003). Liquidity risk and expected stock returns. *Journal of Political economy, 111(3)*, pp. 642-685.

Pedersen, L. H., Fitzgibbons, S., & Pomorski, L. (2020). Responsible investing: The ESG-efficient frontier. *Journal of Financial Economics - Elsevier B. V.*

PRI. (2020). *Principles for Responsible Investment - an Investor Initiative in Partnership with UNEP Initiative and The UN Global Compact.* PRI.

Refinitiv. (2021). *Environmental, Social and Governance (ESG) Scores from Refinitiv* . Refinitiv .

RobecoSAM. (2021). *Robeco - The Investment Engineers* . Retrieved from Smart ESG score: https://www.robeco.com/en/key-strengths/sustainable-investing/glossary/smart-esg-score.html

Russel Investments. (2018). *Materiality Matters: Targeting ESG Issues that can affect performance - the material ESG score.* Russel Investments.

S&P Global. (2018). *GICS*. Retrieved from Global Indusrty Classification Standard: https://www.spglobal.com/marketintelligence/en/documents/112727-gics-mapbook_2018_v3_letter_digitalspreads.pdf

S&P Global. (2020, February 25). *S&P Global*. Retrieved from https://www.spglobal.com/en/research-insights/articles/what-is-the-difference-between-esg-investing-and-socially-responsible-investing

SASB. (2021, April 10). *Sustainability Accounting Standards Board.* Retrieved from Materiality Map - Why is Financial Materiality important?: https://www.sasb.org/standards/materiality-map/

Schaltegger, S., Ludeke.Freund, F., & Hansen, E. G. (2016). Businsess Models for Sustainability: A Co-Evolutionary Analysis of Sustainable Entrepreneurshp, Innovation and Transformation. *Organization & Environment Vol. 29 (3)* , pp. 264-289.

Semenova , N., & Hassel, L. G. (2015, December). On the Validity of Environmental Performance Metrics. *Journal of Business Ethics, Vol. 132, No. 2* , pp. 249-258.

Semenova, N., Hassel, L. G., & Nilsson, H. (2010). *The Value Relevance of Environmental and Social Performance: Evidence from Swedish SIX 300 Companies.* Investment and Corporate Governance Working Papers, Sustainable Investment Research Platform.

Serafeim, G. (2019). *Public Sentiment and the Price of Corporate Sustainability.* Harvard Business School.

Serafeim, G., & Yoon , A. (2021). Stock Price Reactions to ESG News: The Role of ESG Ratings and Disagreement. *Working Paper 21-079.*

Serafeim, G., Khan, M. N., & Yoon, A. (2015, March). Corporate Sustainability: First Evidence on Materiality. *Harvard Business School Working Paper*.

State Street Global Advisors. (2019). *The ESG Data Challenge* . State Street Corporation .

SustainAbility. (2020). *Rate the Raters 2020: Investors Survey and Interviews Results*. ERM Group.

Sustainalytics. (2019). *The ESG Risk Rating - Methodology - Abstract Version 2.0.* Sustainalytics.

Sustainalytics. (2020). *The ESG Risk Rating: Frequently asked Questions for Companies* . Sustainalytics.

US SIF. (2020). *Report on US Sustainable and Impact Investing Trends 2020.* US SIF Foundation.

Wang, Q., Dou, J., & Jia, S. (2015). A Meta-Analytic Review of Corporate Social Responsibility and Corporate Financial Performance: The Moderating Effect of Contextual Factors. *Business & Society* , pp. 1-39.

Woolridge, J. M. (2016). *Introductory Econometrics: A Modern Approach (6th ed.).* South-Western: Cengage Learning.

Zuraida, Z., Houqe, N., & Zijl, T. v. (2015). *Value Relevance of Environmental, Social, and Governance Disclosure.* Wellington: Centre for Accounting, Governance and Taxation Research - Victoria University of Wellington.