Aalto-yliopisto
Aalto-universitetet
Aalto University

Master's Programme in Finance

ESG Rating Disagreement and the Effects on Stock Returns

The Characteristics of ESG Disagreement and its Effects in the U.S. and Europe

---

Iiro Vettervik

Kristian Blitson

Master's thesis

2025

Aalto-yliopisto
Aalto-universitetet
Aalto University

Authors Iiro Vettervik, Kristian Blitson

Title of thesis ESG Rating Disagreement and the Effects on Stock Returns

Programme Master of Science in Economics and Business Administration

Major Finance

Thesis supervisor & advisor Prof. Markku Kaustia

Date 2.6.2025 **Number of pages** 77+4 **Language** English

Abstract

Using data from four prominent ESG rating agencies, we examine the dispersion of ESG ratings and its effects on stock market performance. In a sample of nearly 2000 firms from the past and present index constituents of the S&P 500 and STOXX 600, and with 60,000 monthly firm-rating pairs, we document substantial time-, region-, and industry-persistent disagreement in the ESG ratings. We also explore the implications of this disagreement for stock market returns by studying the performance of various rating dispersion portfolios between 2010 and 2023. We further decompose these portfolios on a second dimension of rating level across total ESG and its component pillars to test whether disagreement has a different effect for sustainable or less sustainable stocks. Despite the seemingly conforming results of the prior research, we fail to replicate their results. Alas, we find investing strategies on ESG rating disagreement rather unattractive. Our study demonstrates — and increases — the lack of research agreement on ESG rating disagreement.

Keywords ESG rating disagreement, ESG ratings, ESG performance, heterogenous beliefs, rating dispersion, stock market returns, uncertainty

Aalto-yliopisto
Aalto-universitetet
Aalto University

Tekijät Iiro Vettervik, Kristian Blitson

Työn nimi Erimielisyys vastuullisuusarvioinneissa ja sen vaikutus pörssituottoihin

Tutkinto Kauppatieteiden maisteri

Koulutusohjelma Rahoitus

Työn valvoja & ohjaaja Prof. Markku Kaustia

Päivämäärä 2.6.2025 **Sivumäärä** 77+4 **Kieli** Englanti

Tiivistelmä

Tutkimme erimielisyyttä vastuullisuusarvioinneissa ja tämän vaikutusta pörssimenestykseen hyödyntäen neljän johtavan ESG-arvioijan dataa. Otoksemme kattaa noin 2000 yritystä ja 60 000 kuukausittaista vastuullisuusarviota S&P 500 ja STOXX 600 -indeksien nykyisistä ja aiemmista yhtiöistä. Havaitsemme systemaattista arvioerimielisyyttä ajasta, alueesta ja toimialasta riippumatta. Selvitämme erimielisyyksien vaikutusta osaketuottoihin tarkastelemalla yritysten ESG-arvioiden hajontaan perustuvien portfolioiden tuottoja vuosina 2010–2023. Lisäksi jaottelemme hajontaportfoliot toisella ulottuvuudella – vastuullisuuden tasolla niin ESG:n kokonaisuuden kuin sen komponenttien osalta – mikä auttaa arvioimaan, vaikuttaako erimielisyys eri tavalla vastuullisiin ja vastuuttomiin osakkeisiin. Kirjallisuuden aiemmista, näennäisesti yhteneväisistä tuloksista huolimatta emme havaitse poikkeavaa riskikorjattua tuottoa vastuullisuuserimielisyyteen pohjautuvilla sijoitusstrategioilla. Niin ikään emme pidä hajontaportfolioita sijoittajille houkuttelevina. Tuloksemme korostavat tutkimuskentän vallitsevaa epäyhdenmukaisuutta ESG-arvioiden erimielisyyttä koskevissa tulkinnoissa.

Avainsanat ESG, ESG-arviot, vastuullisuusarviot, vastuullisuusarvioiden erimielisyys, pörssimenestys, epävarmuus

Table of contents

List of Tables

List of Figures

1 Introduction

Credit ratings proxy a firm's financial risk and the uncertainty surrounding it. While measuring a firm's underlying credit risk would sensibly be up for subjective interpretation, credit rating agencies are practically unanimous in their views: firm-level credit ratings correlate at around 99 % level (Berg et al., 2022). Therefore, a firm credit-rated "Aaa" by Moody's is likely to be rated "AAA" by Standard & Poor's. This could not be further from the truth with sustainability ratings.

Disagreement in ESG ratings has been widely noted during the last decade in research, but predominantly in the U.S. stock markets (see., e.g., Chatterji et al., 2016; Christensen et al., 2022). The correlations between rating agencies have ranged from approximately 75% to as low as 10% depending on the rating dimension and pair (Dimson et al., 2020; Billio et al., 2021; Gibson et al., 2021; Berg et al., 2022). This implies that there exists measurable disagreement — and subsequently uncertainty — on the sustainability of firms[1]. While the causes for the level of this "ESG disagreement" have been recently explored somewhat broadly (see, e.g., Christensen et al., 2022; Berg et al., 2022), the consequences for the disagreement are less so.

In other words, the research is aware ESG rating disagreements exists: *For example, in early 2021, Apple was rated 85/100 by MSCI while 42/100 by LSEG.* Research also has found some explanations for such disagreement: *In case of Apple, MSCI could find the industry in aggregate highly sustainable, while LSEG focuses on Apple's relative performance within*. But lastly, studies are less aware of what this ESG disagreement implies: *What effects on stock market performance would this 43-rating differential have?*

Theoretically, when a company faces undiversifiable, systematic risk, its price should be met with higher expected returns. Differences of opinion in a company's fundamentals, or uncertainty, could be a form of such risk. Disagreement is intuitively the most prone to occur with data that is qualitatively analysed, or likewise, measured on an ordinal scale. One form of commonly used qualitative financial data is analyst forecast estimates. Hypothetically, should a

[1] Literature refers to the differences in ESG ratings of a same firm between agencies with multiple varying terms, including but not limited to ESG rating – "disagreement", "dispersion", "uncertainty", "divergence" or "ambiguity". We mostly refer to the former, often simplified to "ESG disagreement", but employ others for different connotations and contexts. For the purposes of our study, these synonymous terms can be interpreted interchangeably.

company have two differing EPS estimates from two equally reputable organizations, such as 5 and 15, while another otherwise identical one has two ratings of 10, a risk-averse individual should opt for the EPS alternative with less variation, or in other words, less risk. Hence, the price of the latter will increase, which in turn causes its expected returns to lower. This can be applied for sustainability data as well: Should ESG factors be of importance to investors, any disagreement or dispersion can be argued as risk for ESG performance. This would then decrease the perceived value, increasing the company's expected returns.

Heterogenous beliefs in the stock market, such as disagreement, are however also argued to have a completely opposite effect. An idea coined by Miller (1977), which has later gained support in other disagreement research, is illustrated with optimistic and pessimistic market participants. Faced with disagreement, optimistic investors inflate the prices more than pessimists are capable of pushing them down in the presence of short-sale constraints. The higher prices in the short-term would then slowly convert back to the intrinsic value, decreasing the expected returns.

Naturally, understanding this disagreement-return relationship is crucial as ESG factors are increasingly important to investors, yet the inconsistency of ESG ratings could impact asset pricing in ways that are currently underexplored. The few papers we were able to find on specifically the consequences had strict time and geographical confines: solely on the U.S. stock market data with relatively short time periods. Moreover, these papers are lacking uniformity in their results, not just in terms of significance and consistency but also in magnitude and even the sign of the excess returns.

Using data from four major ESG rating agencies of LSEG (formerly Refinitiv and Asset4), MSCI, Bloomberg, and S&P Global, we find a reason to believe the most prominent studies may have overinterpreted the effect, even exaggerated it, to mutually reinforce one another or the proposed theory behind the effect. For instance, Gibson et al. (2021) report total ESG dimension excess returns of 24 basis points (bps) for the 2010–2017 period at the 5% significance level, whereas Avramov et al. (2022) find an insignificant ESG alpha of -2.9 bps for their 2011–2019 period[2]. Nevertheless, both studies make rather strong, generalizing claims on the

[2] As is standard in financial literature, we use the term "alpha" to refer to the intercept coefficient of portfolio regressions on market returns. Specifically, with relation to CAPM, it should would more accurately be "Jensen's alpha", systematized by Jensen (1968). We also review FF3, CAR4 and FF5 alphas, presented later.

observed effect. While differences in methodological approaches may partly explain these discrepancies, the lack of uniformity calls for further inspection of the "dispersion effect" and its premium or discount effect[3].

The purpose of our thesis is to tackle the issue with diverging findings on diverging ESG ratings. To address this gap, we study the implications of ESG disagreement more comprehensively. We hope to contribute clearest on this main topic of interest, the effects of ESG disagreement. We directly broaden the scope of existing studies by extending the time frame, expanding the geographical coverage, increasing the sample size as well as simply revisiting the results. We also bring together the two main papers on the issue to help resolve their conflicting findings. Moreover, our study also complements the general literature of heterogenous beliefs in the stock market, as the findings of this broader research area are also lacking harmony.

We also further test whether the effects of rating disagreement are explained by different levels of ESG, for example, whether effects of disagreement differ between "brown" and "green" firms. To the best of our knowledge, we are the first to sort the disagreement portfolios into a second dimension of the rating level across both total ESG and its pillars. Moreover, to help improve the robustness of our findings and verify the results of prior research, we complement our primary analysis with tests of various methodological approaches. We combine their resulst to examine whether the differences in prior studies can be explained by altering the methodology. We briefly review varying levels of lag in portfolio formation and different proxies of disagreement, such as one controlling for industry variation in the ESG rating dispersion calculations.

Besides the effects of disagreement, we begin our analysis by contributing directly to the general ESG disagreement research. We help understand the vast difference in ratings by reviewing pairwise correlations in both regions as well as the starkly differing shapes of distributions across raters. We also explore differences in levels of disagreement throughout industries. Reviewing general dispersion is especially important in Europe, as we find lacking prior research in the region. The general analysis of ESG rating dispersion also lays the groundwork for the main dispersion effect analysis. Our results for general ESG rating dispersion are harmonious

[3] For brevity and continuity, we hereafter use the term "dispersion effect" and its possible "premium" or "discount" value to refer to the effect of ESG rating disagreement on stock market performance. For instance, if dispersion is to cause undervaluing of assets, it results in dispersion discount.

to those of prior studies — we document and support the existence of considerable disagreement in the ESG ratings. The causes of this disagreement are beyond the scope of our study, but we report the reasons studies have both empirically found and theoretically proposed.

For the stock market implications, contrary to our expectations, we find insufficient evidence that ESG rating dispersion persistently affects stock performance. The European market shows practically no signals of performance differences based on rating dispersion levels, while the U.S. market provides only weak indications. Even in the U.S., the dispersion appears to have an effect in the opposite direction to what we hypothesize: the governance dimension exhibits negative alpha returns across all models during 2010–2015, implying that stocks with low rating dispersion outperform high rating dispersion stocks. This effect, however, disappears in the later period, 2016–2023, during which none of the ESG dimensions produce significant high-low excess returns in either market region. This finding contributes to the understanding of ESG dispersion, suggesting that the effect may be weaker and less consistent than previously thought (e.g., Gibson et al., 2021; Avramov et al., 2022).

Our findings are further supported by the robustness checks, where our results — or lack thereof — persist through different methodological choices. While earlier studies both theorize and report evidence of dispersion premium, our findings suggest that the phenomenon has been transient rather than persistent. A market with green preferences, as discussed by Avramov et al. (2022), could explain why the effect may have offered a premium to investors in the past but has since dissipated: the increasing recognition of the non-monetary benefits of ESG investing partly substitute for the monetary returns. We also present a tentative theoretical explanation to this relationship based on the cross-sectional implications of the two presented contradictory theories on heterogenous beliefs in the stock market.

The remainder of this paper is structured as follows. Section 2 reviews the existing literature on ESG investing and the disagreement among rating agencies. Section 3 outlines our data sources, including geographical coverage, rating agencies, time periods, dataset cleaning, and the rating scale distributions. Our methodology is detailed in Section 4, covering the measurement of ESG rating dispersion, portfolio formation, regression analyses, and the additional tests. Section 5 presents our results, first analysing ESG rating dispersion, then evaluating dispersion portfolios on a second dimension, and finally reviewing the robustness checks. Section 6 summarizes and discusses the implications of our findings. Section 7 briefly concludes.

2 Literature review

This section covers the relevant research on our topic and its related issues. We begin by covering empirical ESG and CSR research concisely and broadly, especially their relationship to financial performance. This is followed by theoretical frameworks pertaining to both ESG studies and stock market disagreement. In the third subsection, we address the prior papers on ESG rating disagreement, with a focus on the few specifically related to its stock market implications. Lastly, as the topic is still highly novel, we briefly present some emerging research papers. Altogether, we aim to contextualize ESG rating disagreement within ESG research and set the stage for our contribution. The relevant research, especially theoretical studies, also helps us formulate the hypothesis for our results.

2.1 ESG Investing

Certainly for decades, ESG matters have been of interest to academia. While the specific term "ESG" was officially adopted in 2004, studies on different aspects of ESG have been published since at least the 1970s (e.g., Preston, 1978). ESG refers to the environmental, social and governance aspects of a company. Essentially, it evaluates the practices of a firm through three different lenses of sustainability. Amidst accelerating climate change and heightened societal awareness, it has undoubtedly ensured a spot as a major fundamental source of information, something that investors either intrinsically care for, or are forced to value for its implications in the stock market. With sustainable investing reaching record high AUMs — such as PRI signatory base peaking last year, yet again, at nearly $130 trillion (PRI, 2024) — ESG research remains in the forefront of financial literature.

Empirical ESG research has largely revolved around its relation to firm value and financial performance. Major meta-analyses and systematic reviews have compiled hundreds of papers on the matter (e.g., Friede et al., 2015; Khan, 2022; Atz et al., 2023). By and large, ESG issues have had an impact on the price performance of stocks. While the evidence remains mixed, even inconclusive, a significant portion of the studies have found a positive relation between the two (Friede et al., 2015, Atz et al., 2023, Chen et al., 2023). Nevertheless, though debated, the sheer number of studies highlights the importance of non-financial factors in financial decision-making.

Some early well-recognized papers linking the closely related term of corporate social responsibility (CSR) to financial performance have also concluded conformingly: CSR and financial performance seem to have a positive, albeit weak, correlation (e.g., Cochran & Wood, 1984; Orlitzky et al., 2003). Once more, this generalization is heavily contested. Some major papers argue the noted relationship is a case of misspecification (McWilliams & Siegel, 2000), the simplicity of traditional statistical techniques (Nelling & Webb, 2009) or measurement issues (Galant & Cadez, 2017). The last on the list, which is a commonly identified cause, is also at the heart of ESG rating disagreement research, a root cause of sort. The relationship between the two could also still be positive, even if not linear, much like Nollet et al. (2016) argue: CSR investments only pay off for financial performance after a threshold investment has been reached. Moreover, CSR may influence financial performance through completely different channels. It can, for instance, mitigate future crash risk (Kim et al., 2014) or soften the blow for the eventual downfall (Lins et al. 2017).

Suffice to say, both ESG and CSR links to firm performance are hardly clear, let alone unanimous. Again, we intuitively suspect a positive correlation between the two. This could stem from multiple different causal channels (see, e.g., Henisz et al., 2019), but also merely due to investor tastes and preferences: stocks favoured by market participants, such as those deemed socially acceptable, would reasonably exhibit higher ex post valuations. Yet, the jury remains out and the evidence mixed. The counter-intuitive relationship — or lack thereof — could be reconciled with disagreement in ESG ratings. The results could have sensibly been less ambiguous had there only been a single, reliable source for ESG information.

Studies have also branched out to explore more than the direct relationship to financial performance. Major academic topics of interest include materiality of specific ESG factors, ESG risks, investor behaviour, regulatory impacts, and rating methodologies. Studies tend to conclude that materiality matters. In other words, not all ESG factors are equally relevant; instead, financially material ESG factors tend to drive better outcomes (Khan et al., 2016). Furthermore, ESG can help identify and mitigate long-term risks, especially governance and environmental risks (Albuquerque et al., 2019). In a similar fashion, institutional investors increasingly integrate ESG due to regulatory, reputational, and stakeholder pressure. Indeed, funds have a significant impact on inflows (outflows) following a categorization of high (low) level in sustainability

(Hartzmark, 2019). Better standardized regulation also shapes the field, but with once more varying effectiveness (Haji et al., 2022).

Research on ESG's effects is premised on the assumption that market participants value ESG in their investment decisions. Firstly, different surveys on ESG investment support that investors care of more than sheer financial performance (see, e.g., Amel-Zadeh & Serafeim, 2018; Riedl & Smeets, 2017; Rau et al., 2024). Even though Amal-Zadeh & Serafeim show that either direct or indirect relevance to financial performance, such as through client demand or product strategy, is the main motivator, ethical considerations play an important role.

Then again, it is theoretically sufficient if the end-user clientele of the investee companies cares for ESG matters, provided investors actually implement ethical practices. Multiple studies have further shown that firms have materially benefited from this positive ESG information (e.g., Cheng et al., 2014; Khan et al., 2016). These were also likely not just attempts of investors to gain improved financial performance through ESG channels, since several studies indeed argue investors are willing to sacrifice financial performance for more sustainable investments (Riedl & Smeets, 2017; Giglio et al., 2025). The study analysing survey data in the U.S. (Giglio et al., 2025) shows that it is not just the large, institutional investors caring for ESG concerns, but instead sustainable investments are made regardless of gender, age, political preferences, and socioeconomic status. The actual ESG investment behaviour throughout these social brackets is also surprisingly similar — conditional on having ESG investments, that is.

The ubiquitous research around sustainability of companies ultimately rests on the reliability of ESG ratings. Investors do not use a single source of ESG information; instead, market participants rely on multiple different data sources for their investment decisions on ESG matters (Hirai & Brady, 2021)[4]. We believe rating agreement between agencies is a strong indicator for reliability, a backbone to it. To even begin questioning the usefulness of ratings, they should be similar across agencies. With enough disagreement, there can be a reason to doubt even the most seminal of papers. Had the data been of different source, would there even have been any results? And by extension, would future rating accordance help reconcile the research discord of the past?

[4] Hirai & Brady further reference a 2020 SquareWell study, which finds vast majority (75%) of asset managers use two or more ESG ratings providers; 40% of them even use four for investment decisions.

2.2 Theoretical Mechanisms

This subsection covers the key theoretical mechanisms relevant to ESG rating disagreement. Ideally, these frameworks — together with prior empirical evidence — help understand what to expect from our analysis, to formulate the hypotheses. First, we illustrate the theorized effects of stock market risk and uncertainty, followed by theories on stock market disagreement. We move on to highlight the foundational theory on ESG investing and lastly, explore the combined two strands for the theory behind ESG rating disagreement.

Firstly, it is important to distinguish between uncertainty and risk, both of which manifest with differing ESG ratings. Stock market disagreements can be viewed through the lens of uncertainty, more specifically Knightian uncertainty (Knight, 1921). Unlike measurable risk, whose outcomes and corresponding probabilities are known, Knightian uncertainty refers to unknown probabilities. When investors have different information or interpret data differently, their opinions on a stock's value diverge. ESG ratings merely proxy for how sustainable a firm is, and the effects of this sustainability are in and of themselves also uncertain. For illustration, it is not only unclear how emissions scores reflect actual emissions, but also what kind of financial and non-financial benefits the efforts of reduced emissions will have. ESG rating divergence therefore further increases the uncertainty about a firm's true sustainability performance and its implications. Whereas higher risk should theoretically always entail higher returns, uncertainty not necessarily. We present two opposing theories on the relationship of uncertainty in the form of disagreement and stock market returns.

2.2.1 Disagreement & Higher Returns

The stock market mantra "higher risk entails higher returns" can, in theory, be applied in the case of disagreement. Market disagreement naturally produces uncertainty for an investor trying to act on the disagreed information. Differences of opinion could also reflect variations in the perceived risk profile, which should then be compensated for in subsequent returns. The standard financial market risk-based theories, such as Sharpe (1964), Lintner (1965), Fama (1970) and Fama & MacBeth (1973) form the basis of this risk-return trade-off assumption. Any disagreements of, for example, analyst forecast estimates or credit ratings, introduce

uncertainty and risk, whose interaction in turn causes lower prices and simultaneously higher expected returns.

The same idea could be extended to the sustainability dimension: Should investors value ESG performance similarly to how they value financial performance, a company whose sustainability is asymmetrically perceived would expectedly face higher future returns. Disagreement in ESG matters would likewise be a risk factor just as those relating to other company fundamentals (see, e.g., Fama & French, 1992; Anderson et al., 2005; Atmaz & Basak, 2018). Then a company with high ESG rating dispersion would expectedly generate higher absolute returns compared to those with lower disagreement.

As is evident from the survey results of the ESG Investing Section 2.1. (e.g., Amel-Zadeh & Serafeim, 2018; Giglio et al., 2025), the former assumption that investors value both ESG and financial performance appears increasingly valid in the stock market. Thus, we find it safe to say that ESG performance plays a significant role for investment decisions. Therefore, any disagreement on ESG performance manifests in uncertainty, which then subsequently causes risk, which, once more, should be compensated for in higher expected returns.

2.2.2 Disagreement & Lower Returns

On the complete contrary however, an idea introduced by Miller (1977), any disagreement or other heterogenous beliefs in the stock market could also entail lower returns. This idea has also been explored later in relevant disagreement studies, such as (e.g., Diether et al., 2002, Sadka & Scherbina, 2007). This opposite relationship theoretically results from the optimistic investors overtaking in importance the pessimistic participants.

Suppose a stock that has investors with either "pessimistic" or "optimistic" views, with an equal number of both. The true, intrinsic price of the stock should be based on their average view. The optimists, perceiving the stock undervalued, are naturally more likely to buy the stock, whereas the pessimists are more prone to sell it short. However, should there exist any short sale constraints or other limits to arbitrage — which there arguably does (see, e.g., Jones & Lamont, 2002; Barberis & Thaler, 2003; Engelberg et al., 2018) — the pessimists are not able to counterweigh the effect of the optimists. The market price then deviates from the true price,

causing short-term overvaluation, and thus subsequent lower expected returns. Essentially, in the short term, optimistic opinions drive up the prices more than pessimistic can drive them down. This effect understandably increases the greater the disagreement becomes.

Once more, the same theory could be extended to the case of ESG disagreement if, again, investors value ESG information in their investment decisions. In this scenario, a stock perceived to be under- or overvalued is simply ESG under- or overrated. ESG optimists, willing to buy ESG underrated stocks, push up the prices more than the ESG pessimists can push them down in the presence of short-sale constraints. Thus, the higher the disagreement in ESG ratings, the lower its expected returns — according to this theory, that is. Stock market uncertainty would now be met with contrary, lower expected returns.

2.2.3 Disagreement & ESG Ratings

We believe that out of these two opposite theories the former is more likely to represent the current market with ESG investing: stocks with higher ESG disagreement should have higher expected returns. Firstly, the latter theory relies on the assumption of short-sale constraints. We believe their effects are decreasing, with markets now fixing the prices more quickly than they used to. If there is a less potent effect of short-sale constraints than a few decades ago, at the time when Miller's optimism theory was introduced, the prices should reflect the optimism less. Financial markets have arguably become more sophisticated, with better infrastructure, more securities available for lending, and some regulatory frameworks that allow for more transparent and accessible short selling. However, even if short-sale constraints are less constraining, the relevant literature seems to remain inconclusive on evidence for more efficient markets (Yen & Lee, 2008; Lim & Brooks, 2011; Martin & Nagel, 2022).

Secondly, while their usage is extremely widespread, we also find the ESG ratings to be relatively unreliable and inaccurate measurements of actual, inherent sustainability. Thus, the gauged ESG rating "disagreement" may not be fully caused by actual difference of opinion, but by nonuniformity of raters. This kind of measurement error would by its nature cause uncertainty and risk to investments, which should be compensated in higher expected returns. This closely relates to the issues on ESG information materiality, which have been argued to cause measurably different results in sustainability research (Khan et al., 2016).

Altogether however, the relationships observed in the market could be a blend of the two. While the implications are contradictory, we do not find the theories themselves mutually exclusive. The interaction of these two opposing ideas could blur the effects observed in the stock market. Essentially, at times short-sale constraints are pushing the price up, while the other times increased risk is rightfully compensated with higher returns. In the cross-section, the effects are then obscured or at least diluted. Then again, should the awareness of disagreement increase, any effect would reasonably weaken, and vice versa.

2.2.4 ESG Investing Theory

Rather than covering multiple papers in detail, we will swiftly present the most important few theoretical papers on sustainable and responsible investing. Nevertheless, most of the well-recognized ones tend to conclude conformingly. In this subsection, we only focus on model-based papers or studies with major theoretical contributions. We review the papers and their models' implications through a lens of ESG rating disagreement.

Comparatively early work by Heinkel et al. (2001) presents an exclusionary ethical investing model, which examines its relationship to cost of capital and thus on investments. Under their model the cost of capital will increase with "unacceptable" firms. A company must care for its footprint even though they do not hold green preferences, if the cost of transforming to "acceptable" operations is lower than the increasing cost of capital resulting from the unacceptable actions. The higher cost of capital is dependent on the percentual share of responsible investors. Fama & French (2007) arrive at similar results discussing how tastes for assets, like those relating to sustainable investments, may compensate lower returns.

A strategic model by Mackey et al. (2007) posits that firms may pursue corporate social responsibility actions not to boost cash flows directly, but to satisfy investor demand. If enough investors value ESG factors, a firm's market value can rise even if profits fall, as investors are willing to pay a premium for "responsible" firms. Investor perception of ESG status would therefore matter for valuation. Similarly, Hong and Kacperczyk (2009) argue that social norms lead to higher required returns for those willing to hold the "sin" stocks. This argument supports that ethical preferences can lead to systematic pricing differences, which supports the idea that ESG-linked characteristics can drive returns independently of fundamentals. Both of these

papers effectively imply that inconsistent sustainability signals, such as those from differing ESG ratings, can lead to mixed investor responses and pricing.

Baker et al. (2022) incorporate a taste-based framework into their model, which then concludes once more that green assets, or in their case particularly bonds, should be priced at a premium. Their model also argues that these assets have more concentrated ownership. Pástor et al. (2021) likewise model investing with ESG criteria. Similarly to the taste-based models, they suppose green stocks have lower expected returns since their investors enjoy holding them. In their follow-up study a year later, Pástor et al. (2022) also believe that the lower returns are partially caused by lower inherent risk caused by greener assets hedging climate risk. Followingly, they price the assets based on both the market portfolio and a "green factor". The positive realizations of this factor may even cause outperformance of green assets. They subsequently argue the factor's effect depends closely on the dispersion of ESG preferences.

Pedersen et al. (2021) include the ESG preferences of investors to their mean-variance extension, "ESG-efficient frontier". The ESG-adjusted CAPM then aims to show not only when greener assets underperform, but also when they outperform. They reduce the optimization problem of financial and sustainable preferences to simple Sharpe ratio – ESG level trade-off; for a given level of ESG, a highest possible Sharpe ratio is determined. Their model implies that considerable increases in ESG levels are only met with minor cuts in the optimal Sharpe ratio. Therefore, it would not cost much to improve on ethical issues, even if the result is still lower returns.

These major studies show that ESG investing relies heavily on investor preferences and perceptions, which affect asset prices, ownership, and cost of capital. By and large, the relationship conforms to the notion of the empirical sustainability studies of Section 2.1: widespread ESG preferences cause comparatively higher valuation, and by extension, lower expected future returns. However, diverging ESG ratings disrupt this mechanism by fragmenting investor views on which firms qualify as responsible. This weakens the pricing effects described in models by, for instance, Heinkel et al. (2001), Mackey et al. (2007), and Pástor et al. (2021), where firm valuation depends on investor consensus. Similarly, the ESG-efficient frontier assumes consistent ESG scores to match preferences with optimal portfolios — again, something rating disagreement complicates.

Avramov et al. (2022) combines the theoretical frameworks behind ESG investing and disagreement. Their paper theorizes that ESG disagreement, or "uncertainty", exposes investors to a market risk in a green neutral environment. In such scenario, a rational investor would likely demand an uncertainty premium to compensate for the additional risk. As a result, ESG dispersion could function as a risk factor that is reflected in market pricing. This theory is therefore harmonious to the first theory behind disagreement, presented in Subsection 2.2.1, in which higher disagreement is met with higher returns. However, Avramov et al. argue the effect becomes more obscure in markets with green preferences, where non-monetary benefits associated with ESG investing may offset the need for a risk premium. Following the reasoning of Pástor et al. (2021), they argue that during an ESG-conscious "green" market sentiment, this preference offsets the dispersion premium. Regardless of the market state, however, disagreement would reduce the demand for ESG assets. In total, across different sentiments, the effects of ESG dispersion on returns would then be inconclusive.

2.3 ESG Disagreement

Ubiquitous disagreement has manifested in a variety of different places, through for example discrepancy in credit ratings and analyst forecast estimates. Although these have been explored broadly since the turn of the century (see, e.g., Cantor & Packer, 1995; Sadka & Scherbina, 2007), research on sustainability rating disagreement has properly emerged only relatively recently. The current research on specifically ESG disagreement can be divided into practically three mutually inclusive areas: the extent, causes, and effects of ESG disagreement.

We start our review from ESG disagreement on a conceptual level. While studies have noted the lack of standardized evaluation of environmental performance, a subcategory of ESG, since the 1990s (e.g., Ilinitch et al., 1998), more thorough and widespread studies on the matter began to appear in the mid-2010s. Still, it was not until the ongoing decade that the topic rapidly gained traction. The arguably most well-known study of ESG rating disagreement, or "divergence", is a 2022 paper by Berg et al., who proxy ESG disagreement with correlations between rating agencies. Firstly, they find ESG ratings correlate on a relatively low level, with pairwise correlations from 6 agencies ranging from 0.40 to 0.70. Correlations for ESG component scores diverge even more strongly, with correlations as low as -0.05 to 0.80. In their sample, both total ESG and component correlations average at around 0.50.

Similarly, Billio et al. (2021) report pairwise ESG correlations between 0.39 and 0.69, although they rely on a lower number of rating agencies. Gibson et al. (2021) find even greater discrepancies, with pairwise ESG component correlations ranging from as low as 0.05 to 0.75. Likewise, Avramov et al. (2022) observe average ESG ratings correlations ranging from 0.25 to 0.70. While these studies use slightly different groups of agencies, they all conclude a high dispersion in ESG ratings.

2.3.1 The Underlying Causes of ESG Disagreement

Intuitively, if ESG information in and of itself is inaccurate, the inaccuracy of ratings will follow. The findings from a study by Chatterji et al. (2009) highlight issues on the accuracy of ESG measurements. They illustrate the ambiguous nature of ESG rating methodologies. Focusing on KLD's ESG rating methodologies[5], the study shows that only around half of the companies with the top 3% of emissions relative to sales were recognized by KLD as producers of substantial emissions. Correspondingly, only half of the most penalized companies relative to sales were labelled as having regulatory problems.

Several other studies have raised concerns about the validity and consistency of ESG ratings as well. Griffin and Mahon, already in 1997, note the long-standing difficulty in assessing corporate sustainability due to inconsistent metrics. Semenova and Hassel (2015) highlight the lack of standardization across widely used proprietary databases. Delmas and Blass (2010) show that ratings vary significantly depending on whether the rating agencies focus on tangible outcomes, like emissions, or softer measures, such as policies and disclosures. This underscores the impact of methodological choices over actual sustainability performance. A later study by Chatterji et al. (2016) further challenges the validity of ESG ratings. They examine the convergence of ESG ratings after considering rater-specific characteristics. Similarly to Berg et al. (2022), they define key contributors to the observed ESG disagreement but with slightly different attributes: lack of common theorization and low commensurability. While the lack of theorization causes every rater to measure by their own definition of "responsibility", the low commensurability contributes to high measuring errors even when assessing similar theoretical

[5] The ESG ratings of KLD (formerly Kinder, Lydenberg & Domini) are now a part of MSCI network and therefore also included in our sample.

concepts. Consequently, the study finds that the rating disagreement does not converge for most raters even when adjusted for methodological differences in measurements, which therefore may indicate invalidity of the ratings on a fundamental level.

Followingly, multiple papers call for uniformity in the ESG metrics. For instance, Billio et al. (2021) show that the lack of standardized ESG metrics and subjective judgments lead to conflicting assessments of the same company. ESG rating disagreement effectively dilutes investor preferences, preventing ESG considerations from meaningfully influencing asset prices. Even when ratings align, the impact on financial performance will remain negligible. They argue that with more compatible metrics, ESG-focused funds could better coordinate investments.

Christensen et al. (2022) specifically examine the effect of ESG disclosure on the level of rating disagreement. They find that greater ESG disclosure increases rating disagreement in environmental and social pillars. Although the relationship may seem counter-intuitive, an increase in information can create more room for varying interpretations, especially in an environment with low theoretical consensus. Essentially, more disclosure leads to more information to disagree about. They also argue disagreement arises more from assessing outcomes rather than input: in the absence of uniform theoretical understanding and frameworks, evaluating the actual implementation and success of policies, such as diversity initiatives, involves considerably more subjectivity than simply noting these policies exist.

Similarly, other studies have attempted to address possible reasons behind the observed high rating dispersion. Berg et al. (2022) find three major causes, or "contributors", for ESG disagreement. Broadly, they divide these three into "scope", "measurement" and "weight". Scope refers to agencies basing their ratings on different "attributes", or in other words, what is included: *One rating agency may include nitrogen emissions while another one may not.* Measurement then refers to how these attributes are measured differently: *Energy consumption can be solely measured by electricity consumption or with different energy sources included.* Weight simply refers to the relative value attributes have toward pillar scores: *CO2 emission grade determines 20 % of the E pillar score.*[6]

[6] These examples are neither from Berg et al. nor do they refer to any specific actions of rating agencies; instead, they are provided for illustrative purposes.

They find that divided into the three, scope contributes around 40% of disagreement. Measurement has the largest effect with 55%, while weight only 5%. On further decomposition of measurement dimensions, they find that 15% of the variation is explained by a "rater effect": a company rated highly on one category is likely to be rated highly by that rater in other categories. Their analysis shows that ESG rating dispersion is not just a matter of definition, but of actual, inherent disagreement. Standardization through regulation would therefore not easily quench the issue on a fundamental level — disagreement may likely just end up manifesting through different channels. For instance, out of the three contributors the one that is arguably the simplest to standardize is weight, the one whose effect happens to also be the smallest. Thus, ESG disagreement will likely persist, and it can subsequently continue to influence asset prices.

2.3.2 The Implications of ESG Disagreement

While several studies have covered the causes of ESG disagreement, the effects and implications of the disagreement have received relatively limited attention in the literature. The first major study examining its consequences emerged only recently, with Gibson et al. (2021) investigating the relationship between ESG disagreement and stock returns. Focusing on S&P 500 stocks, they used rating dispersion from seven leading rating agencies as a proxy for disagreement. They sort portfolios based on the level of rating disagreement and calculate risk-adjusted returns for each strategy. Their findings suggest a significant positive relationship between rating disagreement and stock returns in the environmental (E) and total ESG dimension portfolios between 2010 and 2017. Disagreement in the social (S) pillar is weakly positively related to returns, but its relationship is not statistically significant, while the governance (G) pillar shows no clear effects in either direction.

The other prominent ESG rating disagreement relationship to performance study by Avramov et al. (2022) finds weak evidence for a positive ESG uncertainty alpha over full its 2003–2019 period. However, their subsample analysis shows that this effect fades after 2011, around the time the sample of Gibson et al. (2021) begins. The subsample result therefore somewhat contradicts with their own model's predictions and the findings of Gibson et al. who reports a positive alpha for ESG rating disagreement during the period. Therefore, the results are hardly uniform. This inconsistency — only to be further exacerbated by the lesser-known papers in the topic (e.g., Wang et al., 2024) — motivates the main question of our study.

On a different note, Avramov et al. observe a negative relationship between ESG rating and stock performance among stocks with low ESG uncertainty. This finding follows the previously discussed observations of Pástor et al. (2021), who applied deterministic ESG scores and found a negative CAPM alpha for green stocks. However, Avramov et al. also note that with increasing uncertainty, this ESG-alpha relation becomes nonlinear, which reduces the stock performance predictability for high ESG disagreement stocks.

Another recent study by Serafeim et al. (2023) explores the returns for stocks with high ESG disagreement. The study specifically investigates this through the lenses of sustainability news: how well can ESG scores forecast future ESG news. With rating agreement among the agencies, the scores predict sustainability news well. With respect to stock prices, they identify the ESG rating most strongly correlated with upcoming news and then compare it to the average of other ratings. They find that, in the presence of high ESG rating dispersion, strategies that are following the sentiment of the most predictive rating tend to generate positive excess returns. This result implies that rating disagreement may delay or obscure the incorporation of the most accurate ESG information into stock prices, which in turn should enable excess returns of strategies on ESG disagreement. Their sample, however, is again solely in the U.S., for the total ESG pillar, and during 2010–2018.

Piecing it all together, current ESG rating disagreement studies indicate that low correlations across ESG dimensions largely stem from a lack of rating standardization (e.g., Billio et al., 2021; Berg et al., 2022). Some researchers suggest that even with standardized metrics, disagreements would likely persist (Chatterji et al., 2016; Berg et al., 2022). Furthermore, Christensen et al. (2022) argue that increased ESG disclosure might not resolve the disagreement issue but instead it could even amplify it. For its implications, evidence shows that ESG related disagreement, albeit inconsistently, has recently been associated with positive excess returns in environmental and total ESG dimension. The theory of Avramov et al. (2022) suggests that this effect arises from ESG uncertainty being a market priced risk factor, which warrants a premium. However, they also argue greener market sentiments — essentially times in which markets value sustainability relatively more — could obscure the theorized dispersion effect. Amidst increasing care for ESG matters, this further complicates analysing and interpreting the financial performance of disagreement strategies.

2.4 Emerging Research

As ESG disagreement is still a highly novel issue in literature, we want to briefly highlight some lesser known, emerging research on the topic. We cover evidence from smaller and less developed markets. Similarly, we include the most recent findings. Due to the nature of emerging research and lack of quality assurance, we do not want to overstate their significance but rather present them as ideas for further and new research avenues.

Recent studies have provided some evidence in new markets on the relationship between ESG rating disagreement and stock returns. Vast majority of the latest research revolves around the Chinese stock markets. Contradictory to Gibson et al. (2021), Wang et al. (2024) find a significant negative association between the two variables. They attribute the dispersion discount to reduced investor sentiment, particularly in non-state-owned firms, firms with higher ESG ratings, and those with fewer institutional investors. Zeng et al. (2025) contradict this by instead finding a positive relationship. Meanwhile, Liu et al. (2024) focus on return volatility rather than returns themselves, finding that disagreement increases volatility through noise trading and investor attention. Zeng et al. (2025) also link disagreement to higher idiosyncratic volatility, which they associate with higher stock returns.

Apart from return and volatility implications, recent research has introduced new evidence and perspectives on general ESG disagreement. Much like the return studies, most of the data is again from the Chinese stock markets. The results therefore can understandably have region-specific implications, so it might be difficult to extrapolate them to the western markets. Kimbrough et al. (2024) show that voluntary ESG disclosures can reduce disagreement. This finding contradicts Christensen et al. (2022), as they find that based on a textual analysis, longer ESG reports are related to higher disagreement. A study on Italian firms by Capizzi et al. (2021) presents a framework that decomposes ESG divergence into value and weight components. They identify that weights — especially in social and governance dimensions — are the key drivers for rating divergence. Once more, this is contrary to what Berg et al. (2022) find in their paper, where they believe weight has a rather minute effect.

Another recent study by Luo et al. (2023) argues that disagreement lowers stock price crash risk, particularly in heavily polluting industries. Essentially ESG rating disagreement dilutes

the effect that high ESG ratings generally have on crash risk (e.g., Kim et al., 2014; or for Chinese stock markets, Feng et al., 2022). More recently, Dong et al. (2025) find conversely that stock price crash risk is decreased with greater disagreement, which they attribute to greater media attention achieved with greater disagreement.

Emerging research continues to exacerbate the lack of agreement of relevant literature. Not only are the return findings completely contradictory, but they also show that even the general ESG disagreement research can be conflicting. Although the studies are from different markets and from different periods, it is difficult to argue that rating dispersion would have completely different effect in a less developed stock market region. Seemingly the only constant is the earlier noted effect inconsistency: for example, if environmental rating dispersion causes higher returns, why would social pillar not? These novel research topics and theorizations are nevertheless interesting, which warrants a further exploration also in the West.

3 Data

In this section, we present the data used in our study and provide the rationale behind the key sample choices made. First, we describe the geographical and temporal boundaries and introduce each rating provider used for the rating dispersion calculations. Second, we discuss the characteristics of our dataset and explain the adjustments made to clean and unify the data. Lastly, we illustrate the different rating scales between agencies and detail the distributional properties of the ESG ratings.

3.1 Geographical Coverage

The sample of our study includes ESG rating and return data from the United States and Europe. These markets are represented by the S&P 500 and STOXX Europe 600 indices constituents from 2010 to 2023. We collect ESG ratings from four prominent data vendors for all index constituents during this period. Return data, measured by the monthly total return index (TRI), is obtained from LSEG.

We select the constituents of specifically S&P 500 and STOXX 600 to ensure sufficient coverage and quality of especially the ESG data. These two indices are widely regarded as standard benchmarks, and they offer extensive market reach and diverse sector representation. Both indices also cover most of the floating market capitalization for their respective regions, each with approximately 80–90% of the total. S&P 500 spans the "leading" companies by market capitalization listed in stock exchanges in the United States (S&P Global, 2025) [7]. Similarly, STOXX Europe 600 includes the major exchanges around Europe, not only confined to the Eurozone. It more broadly captures the market by dividing the 600 into segments of small, medium and large free-float market capitalization (STOXX, 2025). Therefore, the comparison of the two indices may not be perfectly appropriate, but we find STOXX 600 to be the best proxy of European stock markets. To accurately replicate the performance of these indices, we adjust the list of constituents monthly; that is, each month has an updated list of constituents[8].

[7] A company is deemed to be included, and characterized as "leading", by a committee. Selection criteria is not only based on the market capitalization, but also for liquidity and volume.

[8] Effectively, monthly update frequency is not necessary for the STOXX 600, which updates its constituents on a quarterly basis, but for simplicity the methodologies are synced.

Essentially, this is done to minimize the effect of potential survivorship bias and to ensure that the data reflects the most up-to-date market conditions. Importantly, it effectively also enables any portfolio strategies from our study to have been implementable in practice.

3.2 Rating Agencies & Time Period

Our four ESG rating providers are LSEG, MSCI, Bloomberg, and S&P Global. These agencies are leaders in the industry and provide comprehensive ESG data (Wong et al., 2019). LSEG and MSCI offer data for the full period from 2010 to 2023, while our Bloomberg and S&P Global samples only span from 2016 onwards. Given the fewer agencies available for the earlier period, we divide our sample into two distinct time frames for a more accurate analysis, while still maintaining our original objective of studying ESG dispersion across the entire 2010–2023 period. The first period with two raters extends from 2010–2015, while the latter with all four covers 2016–2023. The restricted number of rating agencies is discussed in more detail in Section 6.2, Limitations.

Three of our rating agencies, namely LSEG, MSCI, and Bloomberg, are also directly represented in the two studies most relevant to our research, Gibson et al. (2021) and Avramov et al. (2022). However, major consolidations make direct comparison slightly convoluted, as our time frames do not match. LSEG is formerly known as Refinitiv, which in their papers is still referred to as Asset4, its predecessor. Our MSCI scores essentially match their (MSCI) KLD scores, since Kinder, Lydenberg, Domini, acquired earlier by RiskMetrics, later became part of MSCI. MSCI IVA, which both studies use, was later phased out and is since incorporated in the general MSCI scores. S&P Global has acquired Trucost in 2016, then later RobecoSAM, which in turn was used in the sample of Gibson et al. The same goes for the scores of FTSE, now also a member of the LSEG network. Effectively the only major shared provider we are missing in our sample for accurate comparison is Sustainalytics, whom we deliberately chose to exclude from our analysis. They have since fully transitioned to evaluating mere ESG Risk Ratings from previously also including the ESG pillar scores. Gibson et al. also use Inrate's scores, but only for a shorter subperiod.

The decision to begin our review from 2010 primarily stems from data availability. Our MSCI coverage starts in 2007, but the early years were likely influenced by the Great Financial Crisis

(GFC) and the extreme market volatility in its aftermath, which could sensibly distort the impact of ESG factors on stock returns. By 2009–2010, markets were recovering, providing a more stable baseline for analysing ESG related factors. Also, the turn of the decade experienced some advances in sustainable investing and environmental consciousness, such as the Principles of Responsible investing (PRI) having increased sign-ups in the aftermath of GFC and the rapid expansion of ESG investment products and regulatory initiatives. Lastly, the number of rated companies in early ESG data is relatively small for both MSCI and LSEG, which could in turn bias any observed effects due to sample selection issues.

While it is difficult to draw clear-cut boundaries for the periods, early ESG ratings were likely less consistent and had more limited coverage compared to later years, when ESG awareness increased. Intuitively the later we move with ESG data, the higher its quality. The latter time frame indeed coincides with major sustainability improvements like the establishment of UN Sustainable Development Goals in 2015 and the signing of the Paris Agreement later in the same year. While these major advances in sustainability have far from instant effects on specifically the ESG scores, we believe they make later date ratings less comparable to ratings of the earlier period. Moreover, they also ideally influence investors' perception of ESG matters.

3.3 Datasets & Cleaning

The most complete ESG rating data set is for the 2016–2023 period, where all four rating agencies provide ESG scores for the S&P 500 and STOXX 600. While data vendors provide ratings for most companies on a monthly basis, the values are typically updated only once a year. To ensure a more uniform dataset, we require at least three of the four rating providers to have ratings for a company in a given month. This issue is particularly relevant for S&P Global, where ratings data becomes available only later in 2016 and for MSCI, whose coverage extends only until the second quarter of 2023. Moreover, the final year of our review shows noticeably lower index coverage due to its recency, down from over 90% to around 70%. Consequently, we treat 2023 as a dummy variable in our regressions to account for this inconsistency.

To clean our data, we exclude leading and trailing zero values from the raw ESG dataset of each rating agency. These zero scores were highly likely representing either missing data or errors in the reporting process, such as values that failed to be correctly marked as not available,

"N/A". This conclusion is based on the qualitative nature of ESG ratings: even in cases of poor or minimal performance, such ratings are typically reflected in low scores rather than an absolute zero, which would imply a complete absence of measurable ESG activity or engagement. Furthermore, given that the dataset includes only companies from major indices, a value of zero is unexpected. This procedure also serves as a precautionary measure, as these zeros introduced unnecessary noise into the dataset. They also artificially inflate the variation in ESG scores, which may subsequently lead to exaggerated ESG dispersion effects.

Table 1: Rater Agency Characteristics. *Table 1 presents the characteristics of the ESG rater agency data. Time period refers to the time the provider is included in our sample. The number of stocks is the average annual number of stocks included in our monthly samples. The numbers are reported for the US (S&P 500) and EU (STOXX 600) separately, with parentheses values denoting LSEG and MSCI values for the first sample period of 2010–2015. The (C) in dimensions stands for "[ESG] Controversy", which LSEG reports in addition to ESG score, but which we choose to exclude from our sample. Different evaluation methods, abbreviated in the last column, are briefly described below.*

Agency	Time Period	Number of Stocks US / EU		Letter Scale	Numerized Scale	Dimensions	Evaluation Method
LSEG	1/2010 – 12/2023	482 (451)	585 (558)	D- to A+	0-100	ESG (C); E, S, G	Disclosure-focus
MSCI	1/2010 – 6/2023	442 (392)	521 (471)	CCC to AAA	0-10	ESG; E, S, G	Absolute rating
Bloomberg	1/2016 – 12/2023	480	559	None	0-10	ESG; E, S, G	Disclosure-focus
S&P Global	9/2016 – 12/2023	481	541	None	0-100	ESG; E, S, G	Best-in-Class

As displayed in the final column of Table 1, the evaluation methods diverge between agencies. LSEG's evaluation is disclosure-driven, where a company's disclosed ESG data and controversies are compared against sector peers, followed by assigned percentile ranks (LSEG, 2024). MSCI follows an absolute rating method, where it weights companies on 35 key ESG issues and fills any missing disclosures with its own estimates (MSCI, 2024). Bloomberg relies on publicly disclosed data and assesses each company on roughly 30 sector-specific ESG and ESG disclosure issues that are ranked by the probability, magnitude and timing of their financial impact (Bloomberg, 2025; The Good Lobby, 2024)[9]. S&P Global, in turn, combines an annual questionnaire with modelled data and a double-materiality screen before ranking firms against global industry peers (S&P Global, 2024). Altogether, these differing methodologies demonstrate the widely regarded measurement issue of ESG ratings that highlight the lack of standardized ESG metrics.

[9] As Bloomberg does not disclose their exact methodology openly for public use, we add a third-party source for the description.

3.4 Rating Scale & Distribution

Different rating agencies apply different grading scales in the evaluation of ESG performance. These differ in both the scale used and the distribution within. The grading number scales can be rather easily converted to uniform ranges of 0–100, and the data is often already reported as such. The distributions, however, are not as straightforwardly matched. Before combining the ESG datasets of the vendors, the scores should be rescaled and likely standardized to improve comparability. The descriptive statistics of the two datasets are reported in Table 2, while later Figure 1 in Section 5.1 presents the differing distributions visually.

For illustration, the data on MSCI and Bloomberg scores are provided on a scale of 0–10 whereas LSEG and S&P Global use a range of 0–100. We simply scale the former ranges by ten. Yet, this may be hardly sufficient for actual comparison: throughout the current sample (see Table 2), for example, MSCI averages a neatly averaging score of 48, while LSEG has a more benevolent mean score of 63. Simplifying, this would mean a company rated 50 by MSCI should be considered relatively more sustainable than a company rated 60 by LSEG. Naturally, the problem is not limited to differing arithmetic means, but also to different shapes of distribution, be it in levels of standard deviation, skewness, or kurtosis. Therefore, it would be theoretically sound — to a reasonable extent — to normalize these distributions.

Albeit theoretically justifiable, the method of normalization to measure the actual, real-market effects on asset prices requires some limiting assumptions on the "typical" ESG investors. Should ESG rating dispersion be assumed as a systematic risk factor, investors would have to both have access to multiple rating providers and be aware of the differences in the grading distributions. Then they would have to uniformly rescale and redistribute these. Only then could they compare the ratings of two providers meaningfully to gauge the dispersion. Contrarily, if these assumptions do not hold — for example if investors do not standardize the grading scales before comparison — the measured ESG disagreement and uncertainty could also include the differences in distributions of grades.

To make matters worse, the original letter grading applied to specific ranks is also not the same. MSCI reports in total 7 ratings from CCC to AAA while LSEG employs twelve ratings from D- to A+, whereas Bloomberg and S&P Global only assign numerical values. If an investor

only has access to these more qualitative grades, they will have to be aware of this distinction and be able to compare the two. Although a score of, for example, 75 is the lowest A grade for both (A and A-, respectively), MSCI's equivalent is still relatively more difficult to achieve, as the sample mean grade is again, much lower.

Table 2: Descriptive Statistics of ESG Ratings (2016–2023). *Table 2 reports the descriptive statistics of the rating data based on ESG, E, S and G dimensions for the four rating agencies. Below each pillar group, "Total" row in cursive averages the scores for the category. Column "St. Dev." denotes the sample standard deviation. MSCI and Bloomberg scores (originally 0–10) are adjusted to unify the range (0–100) by scaling them by ten. Panel A includes the constituents of S&P 500 for the U.S. data, while Panel B for the European STOXX 600. Both panels only cover our main time frame of 2016–2023.* ***Bolded*** *values highlight the largest value for each column within each dimension grouping.*

Table 2 Panel A: U.S.

Dimension	Agency	Mean	Median	St. Dev.	Skewness	Kurtosis
ESG	LSEG	**60.9**	63.1	16.1	**-0.50**	-0.30
	MSCI	48.8	49.0	9.6	-0.10	0.10
	Bloomberg	37.0	36.2	13.8	0.30	-0.70
	S&P Global	50.4	**50.0**	**28.1**	0.00	**-1.20**
	Total	*49.3*	*49.6*	*16.9*	*-0.1*	*-0.5*
E	LSEG	55.5	58.6	24.1	**-0.40**	-0.80
	MSCI	57.1	55.0	23.8	0.20	-0.70
	Bloomberg	32.4	31.8	19.5	0.30	-0.60
	S&P Global	50.7	**49.0**	**27.7**	0.10	**-1.20**
	Total	*48.9*	*48.6*	*23.8*	*0.1*	*-0.8*
S	LSEG	**62.6**	**64.3**	19.5	-0.30	-0.70
	MSCI	44.9	44.0	15.4	0.20	0.60
	Bloomberg	27.6	22.4	19.1	**1.00**	0.60
	S&P Global	46.2	44.0	**28.9**	0.20	-1.20
	Total	*45.3*	*43.7*	*20.7*	*0.3*	*-0.2*
G	LSEG	64.1	66.7	17.6	-0.70	0.10
	MSCI	53.9	**55.0**	12.7	**-0.60**	0.50
	Bloomberg	**71.1**	72.1	8.2	-1.10	3.50
	S&P Global	56.2	56.0	**25.3**	-0.10	**-1.00**
	Total	*61.3*	*62.5*	*16.0*	*-0.6*	*0.8*

Then again, if no normalization is performed, the level of disagreement and uncertainty could be exacerbated, biased, or even somewhat arbitrary. Each vendor does not have the same share of investors relying on their data. At the same time larger market participants like institutional investors likely do not rely on a single source of rating. Using a simple difference would

therefore likely not result in a perfectly accurate value for the level of dispersion. All of this is to say we examine the effects using both non-standardized as well as standardized ESG ratings to create the portfolios. Non-standardized results will however be only presented in Section 5.3.2 for additional tests. The non-standardized dispersion results simply include scaled but otherwise intact ratings, whereas the standardized portfolios also match the distributions.

Table 2 Panel B: Europe. *See above for Table 2 description.*

Dimension	Agency	Mean	Median	St. Dev.	Skewness	Kurtosis
ESG	LSEG	**63.3**	65.3	16.5	**-0.69**	0.34
	MSCI	55.0	55.0	9.7	0.05	0.16
	Bloomberg	36.7	36.2	13.1	0.21	-0.39
	S&P Global	62.6	**66.0**	**26.9**	-0.40	**-0.92**
	Total	*54.4*	*55.6*	*16.6*	*-0.2*	*-0.2*
E	LSEG	59.8	62.8	22.4	**-0.45**	-0.60
	MSCI	61.0	60.0	23.1	0.09	-0.85
	Bloomberg	33.9	32.7	20.6	0.30	-0.66
	S&P Global	64.4	**68.0**	**25.8**	-0.43	**-0.89**
	Total	*54.8*	*55.9*	*23.0*	*-0.1*	*-0.8*
S	LSEG	**68.4**	**72.6**	19.8	-0.82	0.15
	MSCI	50.0	50.0	16.3	0.19	0.36
	Bloomberg	28.4	24.3	17.7	**0.90**	0.24
	S&P Global	61.9	65.0	**26.4**	-0.34	-0.98
	Total	*52.2*	*53.0*	*20.1*	*0.0*	*-0.1*
G	LSEG	61.6	63.5	21.5	-0.31	-0.84
	MSCI	62.5	**64.0**	14.1	**-0.48**	0.01
	Bloomberg	**62.7**	63.8	13.0	-0.45	-0.18
	S&P Global	60.0	63.0	**28.7**	-0.29	**-1.14**
	Total	*61.7*	*63.6*	*19.3*	*-0.4*	*-0.5*

We standardize the ratings by assigning each stock each month a percentile rating. The procedure closely follows the methodology of, for example, Gibson et al. (2021) and Avramov et al. (2022): We first sort stocks for each month based on their ESG, E, S or G scores by the respective data vendors. For each stock, we determine its percentile within the ratings and assign a corresponding rank. These ranks are then normalized to fall within a 0 to 100 range.

Assigning each stock a percentile rank inherently assumes a flat distribution of the ratings, since every percentile will have the same number of firms. This is therefore fundamentally different to the raw rating provider data, where the shape often, albeit loosely, resembles that of normal

distribution (see Figure 1 for visual distributions). Effectively, in the original data, a rating differential of one unit in the tail ends of the distribution is greater than in the middle. A flat distribution may therefore more closely correspond to how investors actually make their decisions: We believe it's more sensible to assume that a sustainable investor would compare firms in this relative rank way, that is, by simply looking at the relative score placement rather than its absolute value. In other words, if firm A has higher rating than B — irrespective of what's the number value of the difference and where in the distribution the firms lie — firm A is better than B. So even though a flat distribution does not align with the original distribution of the rating providers, we find this an adequate way to standardize the ratings. Nonetheless, the choice of standardization method may impact the results greatly, which is what we aim to test for with our additional tests.

4 Methodology

This section outlines our methodology for the portfolio calculations. It includes description of dispersion calculations, portfolio formation, and regression analyses. Firstly, we compute the ESG rating dispersions for each month using the standardized ESG ratings. We construct a generic portfolio sort with respect to this ESG rating disagreement data. The portfolio's returns are then explained by market factors to measure excess returns.

4.1 ESG Rating Dispersion

As the two sample time periods of 2010–2016 and 2016–2023 have a different number of rating agencies, the calculation of the rating dispersion slightly differs. Since the former period has data from two rating agencies, we proxy for the ESG disagreement with absolute difference of the grades. For example, for a company with MSCI rating of 50 and LSEG rating of 75, the absolute difference, or the level of disagreement, is |50 - 75| = 25. For the latter period, we calculate dispersion based on the sample standard deviation of three or four ratings for each firm in each month. This reflects the average spread of these scores around their means, which we find a robust proxy for the rating disagreement. Both calculation methods are generalized below for clarity.

Formulas I–II: ESG Disagreements. *Formulas I & II define ESG disagreement for a firm i in a month m. Different scores of x, y, z and k are given by different agencies for the same firm. Formula I defines it for the first time period with two scores from two raters, whereas II for the latter period with a minimum of three scores (out of x,y,z and k). STDEV.S stands for sample standard deviation.*

I ESG Disagreement, 2010–2015: $DIS\_ESG_{i,m} = \left|x_{i,m} - y_{i,m}\right|$

II ESG Disagreement, 2016–2023: $DIS_{ESG_{i,m}} = STDEV.S(x_{i,m}, y_{i,m}, z_{i,m}, k_{i,m})$

4.2 Portfolio Formation

We rank all stocks in each month based on the level of dispersion in environmental (E), social (S), and governance (G) pillar, as well as the combined ESG score. The objective is to categorize stocks by their disagreement level and divide them into quintiles. Stocks exhibiting the

highest (lowest) dispersion are allocated into quintiles Q1 (Q5). These quintiles form the basis for our long-short dispersion effect analyses.

To make the strategy investable, we use 6-month lagged portfolio returns — that is, matching ESG dispersion portfolios at month m with the portfolio returns at m + 6. A lag of six months is chosen to alleviate the issue of delayed ratings from rating providers. Some ratings are essentially assigned and added retrospectively to the databases, which distorts the actual information available at a given time. For example, ratings observed today for January 2022 might not have been available to investors until March or April of the same year, at times even later. This delay could make an investment strategy with one-to-three-month lag simply unrealistic. Similarly, it also hampers with ex post research for the dispersions. However, it is likely that some market information from the changed ratings is already partly incorporated into returns before we observe them in the 6-month lag portfolios, which potentially diminishes the significance of any rating dispersion effects. In other words, the longer the selected rolling time window, the greater the stock market noise. We therefore perform robustness tests using various lag periods to assess the magnitude of these timing issues on our findings. We outline the implications of lag in more detail in Section 4.4.

Since the size of the quintiles fluctuates due to variations in ESG rating availability, we adjust the monthly group sizes accordingly. This is particularly relevant in the first sample period of 2010–2016, when ESG dispersion is calculated using only two ratings. In such cases, a single missing value results in the exclusion of a company from the group. For the latter sample period of 2016–2023, a single missing value is not exclusionary as long as a given constituent has at least three other ESG ratings available.

For each formed quintile, we match the constituent stocks with their respective total return index (TRI) data using individual ISIN codes. We rebalance the constituent groups monthly, to ensure only active members of indices are included. The portfolios are then constructed on an equally weighted basis, since we believe this more accurately captures the essence of what we are trying to measure: precisely the average, general effect of ESG rating dispersion — not to overweight specific company level dispersion effects. We construct multiple 6-month lag portfolios, including long-short and long-only strategies with both pillar specific and combined ESG configurations.

4.3 Regression Analyses

The performance of the constructed portfolios is assessed using four widely recognized asset pricing models: the Capital Asset Pricing Model (CAPM), introduced by Sharpe (1964) and Lintner (1965); the Fama-French three-factor model (FF3), proposed by Fama & French (1993); the four-factor model (CAR4), presented by Carhart (1997); and the Fama-French five-factor model (FF5), introduced by Fama & French (2015). The regression specifications are as follows:

Formulas III–VI: Regression components. *R_t is the excess returns of the portfolio at time t, whereas α represents the portfolio's abnormal return, the regression intercept coefficient. Mkt-Rf denotes the market returns excess of risk-free rate. Fama-French three-factor model (FF3) introduces the size Small Minus Big factor (SMB) and quality High Minus Low factor (HML). In the Carhart four-factor model (CAR4), FF3 is extended by adding the momentum factor (MOM). Finally, the Fama-French five-factor model adds the profitability Robust Minus Weak factor (RMW) and the investment Conservative Minus Aggressive factor (CMA).*

III CAPM: $R_t = \alpha + \beta\left(Mkt - R_f\right)_t + \varepsilon_t$

IV FF-3: $R_t = \alpha + \beta_1\left(Mkt - R_f\right)_t + \beta_2\, SMB_t + \beta_3\, HML_t + \varepsilon_t$

V CAR4: $R_t = \alpha + \beta_1\left(Mkt - R_f\right)_t + \beta_2\, SMB_t + \beta_3\, HML_t + \beta_4\, MOM_t + \varepsilon_t$

VI FF-5: $R_t = \alpha + \beta_1\left(Mkt - R_f\right)_t + \beta_2\, SMB_t + \beta_3\, HML_t + \beta_4\, RMW_t + \beta_5\, CMA_t + \varepsilon_t$

For all regressions we conduct Breusch-Pagan and Breusch-Godfrey tests as statistical robustness checks to measure for heteroskedasticity and autocorrelation in our regression models. In cases where these issues are likely present, we apply Newey-West (Newey & West, 1986) standard errors to correct for heteroskedasticity and autocorrelation.

4.4 Additional Tests

To better understand the characteristics of the underlying stocks and their associated rating dispersions, we implement a two-dimensional sorting approach. First, stocks are sorted based on their average ESG dispersion level and divided into three portfolios, “high dispersion”, “low dispersion” and middle dispersion level groups. Within each group, stocks are further

categorized by their average ESG level similarly into three subgroups. For each categorization step, we also compute the high-low spread.

This approach follows a similar concept to that of Avramov et al. (2022), but we adjust the methodology to fit our dataset: we construct nine (3x3) portfolios instead of 25 (5x5) and extend the analysis to each individual ESG component. We believe we are first to perform this on the pillar level. Likewise, to the best of our knowledge, this double sorting has not been performed in the European markets. The adjustment is required for larger portfolio sizes since, in contrast to Avramov et al., we only include index constituents. We hope more observations lead to more robust results within each category. Overall, this test is intended to capture whether the potential dispersion effect differ between "brown" and "green" stocks[10].

In addition to our baseline regressions and the two-dimensional sort on the different factor models, we perform several checks to both test whether our findings are robust for modest changes in methodology and whether the methods of prior literature are explaining the lack of uniformity in the results. These relate to industry-adjusting, dispersion calculation, rolling window basis as well as the non-standardized dispersion tests. Since ESG ratings can vary systematically across industries — as we document in Section 5.1 — we firstly adjust for industry effects by calculating monthly ESG dispersion within industry groups, much like Gibson et al. (2021). Even though we are more interested in the general effect of ESG dispersion rather than industry-specific variations, we want to explore whether the effect persists after accounting for industry differences. The adjustment is made by demeaning the monthly ESG dispersion of a firm by the industry average, as specified in Formula VII.

To test the robustness of our ESG dispersion measure, we compute dispersion using pairwise standard deviation instead of relying solely on the standard deviation. As shown in Formula VIII, the calculation simplifies to scaling each pair difference by a factor of $\sqrt{2}$, which effectively causes slightly less weight for outlier rating values[11]. This might more accurately portray

[10] We use the terms "green" and "brown" hereafter to refer to more sustainable or less sustainable stocks irrespective of the pillar ESG category. That is, a firm with relatively low governance score, for instance, is classified as 'brown' just like a highly polluting, environmentally underperforming company is.

[11] Since our first period of 2010–2015 has only two rating providers, utilizing the pairwise differences will not affect the order of firms in the rankings. In other words, our portfolios would end up with exactly the same constituents. Hence, we only perform the pairwise adjustment for the latter period.

how investors perceive ratings with variation. Essentially, outliers may be even fully disregarded — or at least their effect minimized — at the aggregate, market-wide level. While a smaller change like this should not bear too much weight for the final results as most of the portfolio constituents will remain the same, we hope to partially explain the differences in the two main papers on the topic, Gibson et al. (2021) and Avramov et al. (2022), out of which the latter employs pairwise standard deviations throughout their sample. They justify this approach as it helps compare the dispersions between observations with three and four providers.

Formulas VII–VIII: Varying definitions for disagreements. *Formulas VII and VIII clarify how we define the industry-adjusted ESG dispersion and the pairwise dispersion. Both formulas apply for a firm i in a month m. The latter term in industry adjustment (VII) calculates the average ESG disagreement for an industry, or sector, s. The pairwise form (VIII) is repeated for how many pairs there are for a given firm and then averaged across these. For instance with four ratings, there are six rating agency pairs. These disagreement values are then used as a basis for portfolio formation.*

VII Industry-adj. disp.: $$DIS\_ESG_{i,t,\,\text{ind-adj.}} = DIS\_ESG_{i,t} - \frac{\sum_{i=0}^{n_s} DIS\_ESG_{i,t,s}}{n_s}$$

VIII Pairwise STDEV.S: $$DIS\_ESG_{i,t,\,\text{pairwise}} = \sqrt{\frac{\left(x_{i,t}-\frac{x_{i,t}+y_{i,t}}{2}\right)^2+\left(y_{i,t}-\frac{x_{i,t}+y_{i,t}}{2}\right)^2}{2-1}} = \frac{|x_{i,t}-y_{i,t}|}{\sqrt{2}}$$

Furthermore, we assess the sensitivity of our results to different rolling window lengths when computing ESG dispersion, specifically using 1-month, 3-month, 6-month, and 12-month rolling windows. As outlined earlier in the Section, ESG data becomes available at different times in history, even when it is reported for a specific month ex-post. In other words, an investor in December 2020 might not have access to the ESG ratings that are currently reported available for even July of the same year. Therefore, any investment strategy based on these scores would simply be unimplementable.

To further complicate the issue, the availability of these ratings is also not just for the differences in licenses or general access, but for different providers reporting them differently to history. For instance, MSCI's ratings are updated at the beginning of the month following an ESG rating update, which means even a month rolling basis would be justified. As we only have access to the rating dates of MSCI and LSEG data, we include different rolling time windows for the dispersion calculation.

Naturally, however, increasing the length of the rolling window introduces more noise into the results, hampering with their significance. On the other hand, shorter windows may be uninvestable. For example, Gibson et al. (2021) effectively assume one month rolling window for their January rebalanced portfolios, as they base the portfolios on ESG values of December the previous year. These portfolios are not necessarily purely theoretical, as the ratings will eventually release within their annually updating investment time frame — even if the ESG data is available later, such as in April, any dispersion effects would then be incorporated in the price before the next rebalancing. Nevertheless, the ratings are likely not available already the following month either, which may bias the results. In order to better match their methodology, we also examine the relationship with the January portfolio and simultaneously test whether updating dispersion annually instead of monthly affects the results.

Lastly, to help ensure that differences in rating scaling across providers do not influence our findings, we repeat base tests of the analysis without standardizing ratings across providers. For the non-standardized ratings, we simply rescale the scores of different rating providers to a common range of 0–100, without altering the underlying distributions. Although this method does not reflect a firm's relative ESG performance across different agencies, it might more accurately reflect the raw information available to investors at the time of investment decisions. The rationale for standardization is discussed in greater detail in Section 3.4. All the results of the additional tests are combined in Section 5.3., with specifically robustness checks in 5.3.2.

5 Results & Analysis

In this section, we cover the primary findings and results of our analysis. We begin by analysing the ESG dispersion data. This includes presenting the rating distributions, correlations between providers and the descriptive statistics for the dispersions. We also highlight geographical and industry differences in disagreement levels.

The second part of the section addresses our main research question: the effects of ESG rating disagreement on stock market returns. First, in the second section, we analyse our base regression results and contextualize them with relevant research. The last section focuses on additional tests, where the main topic of interest is a two-dimensional portfolio sorting on both the level of dispersion and the level of ESG. We also assess the robustness of our results and seek to reconcile the varying methodologies of the recent literature.

5.1 Dispersion Analysis & Discussion

Analysing ESG rating dispersion helps understand the extent of disagreement among rating agencies. As most existing research centres around measuring dispersion rather than its impact, we prioritize studying the effects. Examining general dispersion level in our sample remains nonetheless important. Naturally, determining disagreement and its characteristics is the foundation for the dispersion effect studies. Additionally, as noted in literature review, we observe a gap in the general descriptive studies of ESG rating dispersion in specifically Europe. Most of the major studies use specifically U.S. stock market data. Essentially, this section helps us understand whether Europe faces similar dispersions in ratings, and whether there are any major differences between the two, which in turn, could explain any potential differences in implications.

A straightforward and intuitive way to evaluate disagreement in ratings is an analysis of correlations. Table 3 presents the pairwise correlations of ESG ratings across the four agencies. Its coefficients highlight considerable variation. Overall, LSEG with Bloomberg as well as LSEG with S&P Global exhibit the highest correlation pairs, especially for the total ESG ratings, with correlation coefficients of 0.60 and 0.62 in the U.S. and 0.54 and 0.57 in Europe. These coefficients are in line with those documented in earlier literature, such as ones found by Billio et al.

(2021) and Gibson et al. (2021), even among their ranges' higher ends. Yet, even half a decade after these prior samples, correlations remain rather low compared to, for instance, credit ratings. Conversely, S&P Global also happens to show the weakest correlation pair with MSCI with essentially zero correlation. Sensibly each coefficient is positive, so with an increase on the rating of one agency, another is likely to follow — even though at times with extremely weak relation. In both markets, the total ESG pillar has the least variation in correlations, followed by the environmental dimension. This is indeed something we would expect: the aggregate values of ESG, practically by definition, smooth out the differences in the ESG dimensions, whereas environmental data is arguably more objective, with metrics such as carbon emissions or energy use, than issues in social or governance dimension. Similarly, the differences between raters are sensible. On average, the rater agency whose method shares the most qualities with other agencies, LSEG, has the highest average correlation in all its pairs, whereas MSCI, being the only one with direct "absolute rating"-approach, has the weakest.

Table 3: Pairwise Correlations of ESG Ratings Across Agencies (2016–2023). *The table reports the pairwise correlation coefficients of ESG ratings among LSEG, MSCI, Bloomberg (BB), and S&P Global (S&P) for the U.S. (top panel) and Europe (bottom panel). Correlations are computed separately for Total ESG, Environmental, Social, and Governance dimensions. The final column (row) represents the mean, equally weighted correlation for the corresponding row (column). Darker green shading highlights higher correlation values, with filtering performed for both regions separately.*

Region	Dimension	Ref–MSCI	Ref–BB	Ref–S&P	MSCI–BB	MSCI–S&P	BB–S&P	*Average*
US	Total ESG	0.34	0.60	0.62	0.41	0.34	0.53	*0.47*
	Environmental	0.18	0.31	0.56	0.15	0.34	0.39	*0.32*
	Social	0.20	0.37	0.58	0.21	0.22	0.35	*0.32*
	Governance	0.22	0.37	0.23	0.43	0.09	0.26	*0.27*
EU	Total ESG	0.31	0.54	0.57	0.39	0.30	0.45	*0.43*
	Environmental	0.19	0.26	0.50	0.23	0.32	0.34	*0.30*
	Social	0.17	0.32	0.54	0.17	0.16	0.30	*0.28*
	Governance	0.24	0.55	0.30	0.37	0.04	0.25	*0.29*
Total	*Average*	*0.23*	*0.42*	*0.49*	*0.29*	*0.23*	*0.36*	*0.32*

For a long time, prior literature has attributed differences of opinion in ESG matters predominantly to lack of standardization in sustainability metrics (e.g., Griffin and Mahon 1997; Billio et al., 2021). As we document in the Data section, even the mere rating scales are vastly different. Figure 1 illustrates the distribution of ESG ratings across the four providers. This highlights clearly how differently agencies assign scores regardless of the interpretation of the levels.

Figure 1: Distribution of ESG Ratings by Agency (2016–2023). *Figure 1 presents the distribution of ESG ratings for the four rating agencies — LSEG, MSCI, Bloomberg, and S&P Global — across ESG, Environmental, Social, and Governance dimensions in a common sample of S&P 500 and STOXX 600.* ***The x-axis*** *represents the upper boundaries for rating score bins (0–100), and the* ***y-axis*** *shows the frequency of firms within each bin. Panel A displays the distribution for Total ESG scores, while Panels B, C, and D break down the Environmental, Social, and Governance dimensions, respectively. MSCI and Bloomberg scores, originally on a 0–10 scale, have been rescaled to a common 0–100 range for comparability. Similarly, frequencies have been rescaled to match the number of observations of the highest provider (LSEG), in order to better illustrate the differences in distribution shapes. Note: for clarity, the axes are only in the text —* ***x-axis: rating bin; y-axis: frequency of observations.***

Bloomberg's ratings are concentrated in the lower range (20–50), particularly for total ESG and social scores, while S&P Global shows spikes at the upper end. S&P Global's approach is especially evident in the governance dimension. MSCI and LSEG, on the other hand, distribute scores more evenly, with MSCI's grades clustering around an average mean of 50 and LSEG spreading ratings across the scale with slightly more benevolent ratings.

These differences ultimately stem from methodological variations. Bloomberg's lower score follows the disclosure-based approach, which penalizes firms with limited ESG reporting, even if the reported quality is otherwise adequate. While we do not visualize the differences over time, we also found Bloomberg's distribution to have gradually shifted rightward in our annual subsamples, which supports in a way more "objective" measurement; as our data coverage starts already from the aftermath of the Great Financial Crisis, we would expect some gradual improvement over time. Still, as ratings are often compared to one another in the cross-section — that is, at a certain point in time rather than through the years — the more constant shape of LSEG is also defensible. MSCI's clustering around mid-range values matches its normalized "absolute grade" methodology. The differing distributions were presented numerically already in Section 3.4, in Table 2, which additionally highlights the slight differences between Europe and the U.S.

Vast differences in distributions entail disagreement. Even if the data was interpreted precisely the same way, but with different definitions of what constitutes average ESG performance, the ratings will be different enough to capture disagreement. Table 4 presents the descriptive statistics of the disagreement data. It highlights significant disagreement in both regions and across all four dimensions with varying levels.

Table 4: Characteristics of ESG Disagreement Data (2016–2023). *The table reports the descriptive statistics of the ESG rating disagreement data based on ESG, E, S and G dimensions and the ratings of four agencies. "Disagreement" is defined as the sample standard deviation for the available ratings for a given company. MSCI and Bloomberg scores, originally on a 0–10 scale, have been rescaled to a 0–100 range for comparability.*

Dimension	Mean		Median		Standard Deviation		Skewness		Kurtosis	
	US	EU	US	EU	US	EU	US	EU	US	EU
ESG	16.3	17.8	16.4	17.6	5.6	6.1	0.12	0.15	0.06	-0.08
E	21.4	22.1	20.9	21.4	8.9	9.2	0.29	0.30	-0.18	-0.34
S	21.6	23.9	21.5	23.9	7.7	8.3	0.11	-0.01	-0.31	-0.32
G	15.8	15.4	15.2	14.7	6.7	7.3	0.46	0.53	0.08	0.09
Average	*18.8*	*19.8*	*18.5*	*19.4*	*7.2*	*7.7*	*0.25*	*0.24*	*-0.09*	*-0.16*

Social scores exhibit the highest mean dispersion, while governance the lowest, which contradicts the correlation analysis finding that governance ratings are better aligned. The distribution shapes of Figure 1 support better alignment of governance scores — at least in aggregate, that is. In other words, while the rated values cluster close to one another, and thus the average level of disagreement is smaller, the ratings themselves are slightly more randomized. We also find this sensible: The expected performance in governance matters is arguably better than, for instance, in more recently developed environmental issues, which pushes the distributions rightward and closer. Yet, the actual governance measures are difficult to interpret, which makes the scores more arbitrary in the right end. Combined, this results exactly in the observed pattern: correlation is poor, but the scores are merely generally closer aligned, which artificially lowers the disagreement. In the environmental dimension this relationship is reversed with the highest standard deviation but also greatest correlation. The differences in the level of kurtosis help explain this difference in the relationship: with greater kurtosis, governance disagreement is more concentrated, while environmental and social scores have flatter, more dispersed distributions. Overall, ESG dispersion remains persistent across regions and categories. It should be noted that the differences in the disagreement data are, however, relatively minute.

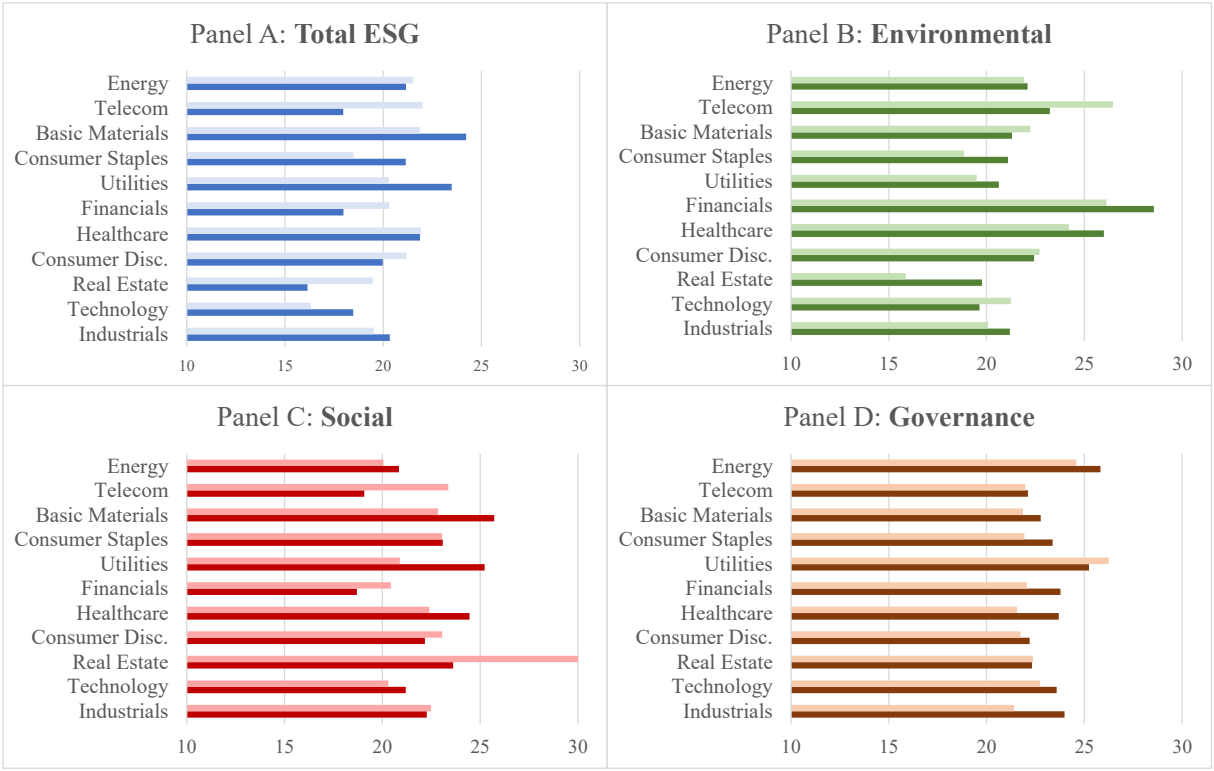
Finally, prior studies (e.g., Gibson et al., 2021) also note the level of disagreement has varied across industries, at least in the U.S. This, once more, is intuitive: ESG information in certain sectors is arguably more comprehendible and transparent, and thus easier to agree on. The direct environmental data about a manufacturing company ("Scope 1") can be easier to quantify than that of, for instance, the financial sector, whose sustainability impacts are mostly more indirect ("Scope 3"). Figure 2 displays the ESG rating dispersion in our sample by industry, including levels in both markets. Indeed, our evidence supports the findings of literature, even if they proxy the disagreement with correlation differences. We choose to highlight them by a level of disagreement, as these values are also employed in the industry-adjusted portfolios.

The dispersions are not glaringly different based on the sector, as most of the values lie within a 15 to 25 range. Yet, the differences in distributions shapes and the nature of disagreement data itself command a greater lower bound but also limit the extreme variation: highly different shapes induce a baseline level of variation, 15 in this case, while four differing rating agencies will gap the upside — if they were to randomly assign the scores, even 25 standard deviation is

a rather unexpected result. In other words, a mere range of ten (between 15-25) supports meaningful industry variation.

Financials, telecom and healthcare show higher dispersion in the environmental scores, whereas real estate, utilities and industrials have lower — consistent with the earlier theorization. Social pillar dispersion, on the other hand, is greater in real estate and basic materials, but lower in financials. Governance dispersion is, much like in the general disagreement data, relatively uniform across sectors. While the regional differences are not particularly stark, or the patterns between dimensions overly consistent, the table reinforces the broader ESG rating divergence issue. These help show that sector-specific factors contribute to disagreement among rating agencies, and by extension, they call for the industry-adjusted ESG disagreement tests. Nevertheless, the patterns below could also be partially due to random noise.

Figure 2: ESG Rating Dispersion Across Industries (2016–2023). *Figure 2 illustrates the dispersion of ESG ratings across industries for U.S. (darker bars) and European (lighter bars) firms. Panels A, B, C, and D depict dispersion in Total ESG, Environmental, Social, and Governance ratings, respectively. The x-axis represents the level of rating disagreement among agencies, while the y-axis lists the 11 FTSE industries. Note: for clarity, axes are only in the text —* ***x-axis: disagreement level; y-axis: frequency of observations.***



5.2 Portfolio Performance

Next, we analyse the performance of ESG rating disagreement portfolios of S&P 500 and STOXX 600 constituents over two distinct time periods, 2010–2015 and 2016–2023. Specifically, we investigate whether the theorized additional uncertainty induced by high dispersion in ESG, environmental, social, or governance components translates to additional risk, compensated by abnormal returns. To address this question, we run regressions of returns on well-known risk factors for each component and portfolio variant. Since we employ different methodological approaches and datasets for the two periods, we present our analysis in two separate parts. We also report tables for factor loadings of the 2016–2023 regressions but only present them in the Appendix (Table A10).

5.2.1 Time Period 2010–2015

Table 5 presents the results for our baseline regressions in the first time period. Beginning with S&P 500 stocks for the total ESG dimension (see Table 5, Panel A), the high-low portfolio yields near zero alphas for all models ranging from -6 to 2 basis points (bps). Since none of the models produce statistically significant intercept coefficients, we find no distinguishable return differences between these strategies. The environmental high-low strategy indicates more positive excess returns ranging from 21 to 30 bps: while the five-factor model has an alpha of 30 bps significant at the 10% level, the overall evidence remains inconclusive. The results on E dimension are, however, within the same magnitude as those of Gibson et al. (2021), who document monthly excess returns of approximately 20 bps during a similar time period. As their coefficients for S and G portfolios are essentially zero, we believe the effect observed in their ESG dimension is simply driven by the E dimension.

The low disagreement portfolios in the social dimension deliver positive and statistically significant alphas at the 5% level in most models, ranging from 18 to 23 bps. Subsequently, the high-low strategy returns negative excess returns but is lacking significance to suggest any distinct advantage in favour of the low leg. The results with the highest significance emerge from the governance portfolios, where the low portfolio consistently produces positive alpha across all models, ranging from 22 to 29 bps, with significance levels reaching the 1% mark in the five-factor model. This relationship remains for the high-low strategy, which yields consistently

negative and significant returns across all models, with five-factor model alpha of -33 bps at the 1% significance level. Interestingly, while specifically the governance ratings had the lowest correlations between all four agencies, the pillar correlation between LSEG and MSCI ratings was in fact the highest in governance. Even if the difference is not substantial, this may influence the results here.

Table 5: ESG Disagreement on Stock Returns 2010–2015. *The table displays equally weighted excess returns based on ESG rating disagreement. The ESG rating disagreement is reported in terms of the Total (ESG), Environmental (E), Social (S), and Governance (G) dimensions. These values are defined as the absolute difference between the pillar scores of two providers, MSCI and LSEG. Columns 3–6 adjust the returns with different pricing models, either CAPM, FF3, CAR4, or FF5. The grey-shaded values highlight the intercept coefficients, or the excess returns, in the different factor model specifications, whereas the white-shaded values beneath them indicate the respective t-statistics. The portfolio "High" represents the highest disagreement, while "Low" the smallest. The "High-Low" row denotes a strategy with a long position on high ESG disagreement stocks and a short position on low ESG disagreement stocks. The "Sharpe" column denotes the Sharpe ratios of portfolios, while "N" denotes the number of monthly time-series observations in each portfolio. T-statistics are adjusted with the Newey-West corrected standard errors, with values in parentheses. "*", "**", "***" denote statistical significance at levels 10%, 5%, and 1%, respectively.*

Table 5 Panel A: U.S.

Dimension	Portfolio	CAPM	FF3	CAR4	FF5	Sharpe	N
ESG	High	0.13	0.16	0.16	0.12	1.13	72
		(1.39)	(1.58)	(1.63)	(1.21)		
	Low	0.12	0.15	**0.21***	0.18	1.11	72
		(1.05)	(1.34)	(1.96)	(1.56)		
	High-Low	0.02	0.00	-0.04	-0.06	-0.13	72
		(0.11)	(0.02)	(-0.29)	(-0.43)		
E	High	0.12	0.17	0.19	0.17	1.10	72
		(1.02)	(1.44)	(1.62)	(1.42)		
	Low	-0.10	-0.06	-0.02	-0.13	0.91	72
		(-0.87)	(-0.47)	(-0.18)	(-1.26)		
	High-Low	0.22	0.22	0.21	**0.30***	0.65	72
		(1.38)	(1.35)	(1.24)	(1.92)		
S	High	0.04	0.09	0.10	0.02	1.03	72
		(0.30)	(0.70)	(0.80)	(0.21)		
	Low	**0.18***	**0.22****	**0.23****	**0.23****	1.17	72
		(1.84)	(2.27)	(2.23)	(2.10)		
	High-Low	-0.14	-0.14	-0.13	-0.21	-0.12	72
		(-0.91)	(-0.86)	(-0.79)	(-1.33)		
G	High	-0.04	-0.01	0.04	-0.04	0.97	72
		(-0.37)	(-0.04)	(0.32)	(-0.39)		
	Low	**0.22***	**0.28****	**0.29****	**0.29*****	1.20	72
		(1.87)	(2.46)	(2.61)	(2.69)		
	High-Low	**-0.26****	**-0.28****	**-0.25****	**-0.33*****	-0.72	72
		(-2.16)	(-2.34)	(-2.11)	(-2.74)		

While governance is the only dimension showing a distinct difference between high- and low-portfolio performance within the S&P 500, the direction is opposite to expectations or the prior results by Gibson et al. (2021). Low dispersion stocks in G significantly outperform their high dispersion counterparts, effectively reversing the intended high-low strategy. This is also visually evident in Panel A of a later Figure 3, in Subsection 5.3.1, where the curve for the G dimension slopes downward before the marked split point of our sample.

Table 5 Panel B: Europe. *See above for Table 5 description.*

Dimension	Portfolio	CAPM	FF3	CAR4	FF5	Sharpe	N
ESG	High	0.33	**0.42***	0.39	**0.61****	0.61	73
		(1.53)	(1.90)	(1.65)	(2.57)		
	Low	**0.54*****	**0.55*****	**0.53*****	**0.63*****	0.86	73
		(2.94)	(2.95)	(2.70)	(3.08)		
	High-Low	-0.21	-0.12	-0.14	-0.02	-0.45	73
		(-1.63)	(-1.01)	(-1.10)	(-0.12)		
E	High	0.30	**0.40****	**0.48****	**0.62*****	0.58	73
		(1.46)	(2.30)	(2.01)	(2.91)		
	Low	**0.50****	**0.44****	**0.44****	**0.59*****	0.80	73
		(2.62)	(2.50)	(2.09)	(2.74)		
	High-Low	-0.21	-0.04	0.04	0.03	-0.35	73
		(-1.49)	(-0.28)	(0.27)	(0.23)		
S	High	0.26	0.35	0.35	**0.44***	0.56	73
		(1.18)	(1.56)	(1.47)	(1.76)		
	Low	**0.38***	**0.37***	0.39	**0.57****	0.67	73
		(1.88)	(1.75)	(1.51)	(2.52)		
	High-Low	-0.12	-0.02	-0.04	-0.13	-0.33	73
		(-0.79)	(-0.13)	(-0.23)	(-0.79)		
G	High	**0.39***	**0.50****	**0.48****	**0.71*****	0.66	73
		(1.88)	(2.62)	(2.11)	(3.16)		
	Low	**0.36***	**0.35***	**0.37***	**0.45****	0.67	73
		(1.85)	(1.69)	(1.72)	(2.03)		
	High-Low	0.03	0.15	0.11	**0.26***	0.23	73
		(0.23)	(1.15)	(0.76)	(1.83)		

The same dispersion effects do not translate to the European stock markets, as shown by the figures in Panel B of Table 5. Overall, across the dimensions, both high and low legs are generating excess returns in tandem, which results in an insignificant long-short portfolio: The ESG low dispersion portfolio consistently delivers positive and significant alpha across all models, ranging from 53 to 63 bps and significant at the 1% level. While the lower disagreement leg yielding more consistent excess returns is opposite to our expectations, the high-low strategy remains statistically insignificant. Similarly for the environmental dimension, both the high and low portfolios produce variably significant and positive alphas. Neither extends to the high-low strategy, where the results from -21 to 4 bps are statistically insignificant. While the social dimension portfolios are mostly significant for the low leg, the difference between the two is, yet again, minimal: −13 to −2 bps. The governance dimension does return a significant alpha,

but only for the five-factor model (26 bps at the 10% level). This relationship is however exactly opposite to that of the S&P 500 governance dimension.

The STOXX 600 results indicate that while high and low disagreement portfolios can generate both negative and positive, occasionally significant alphas, there is no evidence that high ESG disagreement, or the pillars of E, S, or G, systematically outperforms — or underperforms — low disagreement in European markets. Alas, not only does there appear no persistent patterns in the excess returns across ESG dimensions, but the results across the regions are even less consistent.

5.2.2 Time Period 2016–2023

Next, we review the effect of rating dispersion over the main sample 2016–2023 period. Beginning again with S&P 500 stocks, presented in Panel A of Table 6, we observe a general lack of statistical significance across all ESG dimensions and regression models. For the total ESG dimension, there is no meaningful high-low return differential in any of the models, as highlighted by the near-zero spread. The E and S dimensions exhibit only minor return differences that are insignificant across all models. For the G portfolios, the high-low strategy produces positive values across all models (ranging from 12 to 15 bps), but once again, none reach statistical significance. These result contrasts directionally with the previous period in the S&P 500, where governance disagreement had a stronger but opposite effect. This indicates a case of mean-reversion, as is visually evident from Panel A of Figure 3 — the arbitrary outperformance of the past is simply reversed.

For the STOXX 600 during the 2016–2023 period, in Panel B of Table 6, the high-low strategies across all ESG dimensions remain statistically insignificant. This indicates there does not exist a dispersion effect in the European market either. The high-low strategies of total ESG, environmental and governance dimensions all hover around zero. In the social dimension portfolios, alphas for the high and low disagreement portfolios closely mirror each other in both direction and magnitude, which nets out at near-zero. In general, the “excess returns” visible across dimensions in the CAR4 column indicate the failure of the model itself in European market rather than signal of actual alphas. Indeed, this illustrates the joint test problem: every regression on

pricing models simultaneously tests whether there are excess returns, but also whether the model itself is valid.

Table 6: ESG Disagreement on Stock Returns 2016–2023. *The table displays equally weighted excess returns based on ESG rating disagreement. The ESG rating disagreement is reported in terms of the Total (ESG), Environmental (E), Social (S), and Governance (G) dimensions. These values are defined as the sample standard deviation from the pillar scores of four providers: MSCI, LSEG, Bloomberg and S&P Global. Columns 3–6 adjust the returns with different pricing models, either CAPM, FF3, CAR4, or FF5. The grey-shaded values highlight the intercept coefficients, or the excess returns, in the different factor model specifications, whereas the white-shaded values beneath them indicate the respective t-statistics. The portfolio "High" represents the highest disagreement, while "Low" the smallest. The "High-Low" row denotes a strategy with a long position on high ESG disagreement stocks and a short position on low ESG disagreement stocks. The "Sharpe" column denotes the Sharpe ratios of portfolios, while "N" denotes the number of observations in each portfolio. T-statistics are adjusted with the Newey-West corrected standard errors, with values in parentheses. "*", "**", "***" denote statistical significance at levels 10%, 5%, and 1%, respectively. The factor loadings for the FF5 model factors are presented, for illustration, in the Appendix, Table A10. The appendix table also includes the adjusted R^2 values.*

Table 6 Panel A: U.S.

Dimension	Portfolio	CAPM	FF3	CAR4	FF5	Sharpe	N
ESG	High	-0.07	-0.05	0.03	-0.17	0.59	91
		(-0.35)	(-0.32)	(0.18)	(-1.22)		
	Low	-0.08	-0.05	-0.02	-0.05	0.56	91
		(-0.39)	(-0.37)	(-0.10)	(-0.36)		
	High-Low	0.01	0.00	0.04	-0.12	0.08	91
		(0.06)	(0.01)	(0.24)	(-0.67)		
E	High	0.01	0.04	0.08	0.03	0.61	91
		(0.06)	(0.31)	(0.60)	(0.21)		
	Low	-0.02	-0.02	0.04	-0.07	0.62	91
		(-0.10)	(-0.13)	(0.27)	(-0.42)		
	High-Low	0.03	0.06	0.03	0.10	0.08	91
		(0.22)	(0.36)	(0.22)	(0.60)		
S	High	-0.06	-0.05	0.02	-0.12	0.60	91
		(-0.34)	(-0.32)	(0.10)	(-0.94)		
	Low	0.01	0.04	0.08	0.02	0.59	91
		(0.03)	(0.33)	(0.72)	(0.16)		
	High-Low	-0.07	-0.09	-0.07	-0.14	-0.03	91
		(-0.38)	(-0.67)	(-0.48)	(-1.00)		
G	High	0.11	0.12	0.17	0.06	0.66	91
		(0.58)	(1.01)	(1.57)	(0.51)		
	Low	-0.04	-0.02	0.04	-0.06	0.57	91
		(-0.24)	(-0.20)	(0.32)	(-0.54)		
	High-Low	0.15	0.14	0.13	0.12	0.31	91
		(1.32)	(1.27)	(1.14)	(1.14)		

The 2016–2023 period provides even less evidence of non-zero alphas than the 2010–2015 period. While the earlier period indicates some slight dispersion related differences in performance, the later period either shows no persistence or even reverses the earlier findings. This is especially evident in the U.S. governance dimension. Ultimately, the lack of evidence

suggests that rating disagreement does not lead to return differences over time. This, in turn, calls into question whether a dispersion effect exists at all, let alone whether it transforms into a profitable investment strategy.

Table 6 Panel B: Europe. *See above for Table 6 description.*

Dimension	Portfolio	CAPM	FF3	CAR4	FF5	Sharpe	N
ESG	High	0.06	0.06	0.27	-0.03	0.41	91
		(0.28)	(0.30)	(1.45)	(-0.14)		
	Low	0.08	0.09	0.32	-0.01	0.44	91
		(0.42)	(0.45)	(1.63)	(-0.06)		
	High-Low	-0.02	-0.03	-0.04	-0.01	-0.13	91
		(-0.15)	(-0.32)	(-0.41)	(-0.13)		
E	High	0.06	0.06	0.25	-0.01	0.41	91
		(0.26)	(0.31)	(1.21)	(-0.03)		
	Low	0.06	0.07	**0.33***	-0.09	0.41	91
		(0.35)	(0.34)	(1.90)	(-0.40)		
	High-Low	0.00	0.00	-0.08	0.09	-0.01	91
		(-0.02)	(-0.03)	(-0.60)	(0.57)		
S	High	0.17	0.17	**0.37***	0.06	0.45	91
		(0.83)	(0.95)	(1.92)	(0.29)		
	Low	0.17	0.18	**0.42****	0.06	0.49	91
		(0.79)	(0.85)	(2.11)	(0.31)		
	High-Low	0.00	-0.01	-0.05	-0.01	-0.30	91
		(-0.02)	(-0.05)	(-0.40)	(-0.06)		
G	High	0.04	0.04	0.26	-0.07	0.42	91
		(0.19)	(0.21)	(1.37)	(-0.31)		
	Low	0.09	0.10	0.31	-0.05	0.43	91
		(0.44)	(0.49)	(1.49)	(-0.26)		
	High-Low	-0.05	-0.06	-0.05	-0.01	-0.04	91
		(-0.34)	(-0.50)	(-0.43)	(-0.09)		

5.3 Additional Tests' Results

In this section, we review the results of additional tests and robustness checks. We start by exploring the two-dimensional portfolios to investigate whether the dispersion has different effects for "brown" or "green" stocks. In the latter section, we briefly measure the robustness of our baseline results through lenses of three different categories: dispersion calculation methodology, industry-adjustment and the lag period length.

5.3.1 Two-Dimensional Sort

The results of our two-dimensional portfolio are presented in Table 7. Following the 3x3 double sort, the table highlights the high and low portfolios, as well as their long-short spread. The main portfolios of interest, the right column of each pillar section, highlight how different levels of ESG affect the dispersion portfolios. For instance, the top right portfolio, high-low ESG in

high dispersion, effectively indicates whether the high scoring ESG stocks outperform low scoring ESG stocks within the high dispersion group. These top-right portfolios are visualized in Panels C and D of Figure 3. The middle-right portfolio tests the same but within low dispersion setting. The double-sorted tests are still performed on a relatively surface level, and, hence, are not our main topic of interest. Yet, they help provide a general idea whether disagreement has different effect for sustainable or less sustainable stocks for each dimension. Altogether, there does not seem to be strong or persistent effect across markets or dimensions, even if there are interesting singular observations.

Specifically, the bottom right corner in each pillar section, or the double high-low portfolio, helps make the distinction — whether dispersion effect is different between ESG levels. The portfolio is calculated from the difference in the high-low dispersion portfolios between high and low score ratings. Essentially, it follows the difference in the bottom row portfolios: the combined portfolio takes a long position in the bottom left portfolio and a short position on the bottom middle portfolio. Therefore, it is more of a theoretical approach, but ideally, the portfolio captures the difference between brown and green stocks in the dispersion effect: If it is significantly positive (negative), green stocks are relatively outperforming (underperforming) brown stocks within the dispersion effect. If there is no difference, the effect is equally strong — or there is no effect at all. To implement this in practical terms, an investor would buy a combination of (High Disp., High ESG) and (Low Disp., Low ESG) stocks, while shorting (High Disp., Low ESG) and (Low Disp., High ESG) stocks. We mostly focus our analysis on these portfolios as we believe they can concisely capture what we are mostly interested in. All intercept coefficients come from the Fama-French five-factor model regressions.

In the U.S. market between 2010–2015, presented in Panel A of Table 7, the double-sorted high-low portfolio has a highly significant negative alpha for the governance pillar. The portfolio delivers negative excess returns of -70 bps at the 1% confidence level, which, in the high-low dispersion setting, implies that unsustainable stocks are relatively outperforming in the dispersion setting. While the result is primarily driven by the bottom left portfolio — the underperformance of high governance stocks within the dispersion effect — the results essentially argue that in the presence of disagreement, less sustainable governance stocks should be favoured. For the other pillars, with alphas ranging from -17 to 11 bps, the sustainability characteristics seem not to affect the dispersion effect as clearly.

Table 7: Two-Dimensional Sorting by Average ESG Rating and Rating Dispersion Level 2010–2015*. The table presents equally weighted FF5 model excess returns of portfolios sorted twice on ESG rating dispersion levels and ESG levels in the 2010–2015 period. The stocks are first sorted into quintiles based on ESG or its pillar rating dispersion, followed by a secondary sorting based on the average rating level of the respective component. Out of the nine portfolios (3x3), for brevity, we report only the results for the top (High) and bottom (Low) categories, as well as the long-short spread, each for both dimensions. Combined with the double High-Low-portfolio in the bottom right corners, this amounts to nine portfolios for each of the four pillar groups. All the numerical values are the FF5 model intercept coefficients and their t-statistics. Columns represent the ESG level, while rows indicate the level of rating dispersion based on the sample standard deviation of ratings from at least three providers out of MSCI, LSEG, Bloomberg, S&P Global. For instance, the cell labelled "High ESG, Low Disp." reports the excess returns of the highest ESG rating tercile within the low dispersion dimension tercile. T-statistics are adjusted with the Newey-West corrected standard errors, with values in parentheses. "*", "**", "***" denote statistical significance at levels 10%, 5%, and 1%, respectively. The interpretation of the table is explained in greater detail in the beginning of the section.*

Table 7 Panel A: U.S.

S&P 500	ESG			E			S			G		
	High	Low	High-Low	High	Low	High-Low	High	Low	High-Low	High	Low	High-Low
High Disp.	0.13	**0.33****	-0.20	-0.02	-0.25	0.22	0.13	0.23	-0.10	-0.24	**0.30****	**-0.54****
	(1.23)	(2.25)	(-1.16)	(0.30)	(1.55)	(-1.33)	(1.13)	(1.49)	(-0.60)	(-1.28)	(2.41)	(-2.42)
Low Disp.	0.10	0.14	-0.04	-0.02	-0.13	0.11	0.01	0.12	-0.12	**0.37*****	0.20	0.16
	(0.65)	(0.78)	(-0.16)	(-0.17)	(-0.83)	(0.53)	(0.04)	(0.77)	(-0.64)	(2.79)	(1.49)	(0.87)
High-Low Disp.	0.03	0.19	-0.17	0.00	-0.11	0.11	0.12	0.11	0.01	**-0.60*****	0.10	**-0.70*****
	(0.15)	(0.83)	(-0.60)	(-0.01)	(-0.47)	(0.41)	(0.72)	(0.67)	(0.05)	(-3.20)	(0.55)	(-2.65)

Table 7 Panel B: Europe.

STOXX 600	ESG			E			S			G		
	High	Low	High-Low	High	Low	High-Low	High	Low	High-Low	High	Low	High-Low
High Disp.	**0.57***	0.40	0.17	-0.42	-0.33	-0.09	**0.70****	0.29	**0.41****	**0.85*****	**0.87*****	-0.02
	(1.80)	(1.43)	(0.81)	(1.47)	(1.18)	(0.35)	(2.59)	(1.05)	(2.18)	(3.43)	(3.47)	(-0.09)
Low Disp.	**0.46****	**0.51***	-0.05	**0.64****	**0.58****	0.06	0.40	**0.46***	-0.06	0.41	**0.41***	0.00
	(2.06)	(1.98)	(-0.23)	(2.51)	(2.38)	(0.23)	(1.58)	(1.82)	(-0.31)	(1.42)	(1.68)	(-0.01)
High-Low Disp.	0.11	-0.11	0.22	**-1.06****	**-0.91***	-0.15	0.30	-0.18	**0.47***	**0.44***	**0.46****	-0.02
	(0.58)	(-0.52)	(0.90)	(-2.11)	(-1.93)	(-0.40)	(1.49)	(-0.88)	(1.96)	(1.75)	(2.49)	(-0.06)

The results are different in the European market, presented in Panel B. The statistically significant alphas of environmental low dispersion stocks, with 64 and 58 bps, indicate that in Europe,

low dispersion environmental stocks have outperformed regardless of whether these stocks are in the high or low tercile of environmental level. This is further highlighted in the bottom row: the E coefficients of high-low dispersion portfolios imply that irrespective of environmental rating level, high dispersion stock portfolios have relatively underperformed. This relationship is reversed for the governance portfolios: irrespective of governance rating level, high dispersion appears to outperform low dispersion. Importantly both of these significant results seemingly contradict with the prior findings of European dispersion effect in Section 5.2 (see Panel B of Table 6). Essentially, the average value of the two bottom row values — -1.06 and -0.91 in E, while 0.44 and 0.46 in G — could be expected to converge toward the value observed without the second sort on ESG level. Indeed, these averages do not match: the FF5 intercepts were originally 0.03 in E, while 0.26 in G. Averaging across high and low tercile however omits the middle portfolio, the remaining tercile between the high and low legs. Still, it can be difficult to argue that the average level ESG scores would have such a different dispersion effect. The other difference between the two methodologies of different portfolio sorts could also explain the results: one is divided into 3x3 portfolios, the other into 5x5. Less optimistically, however, it could be due to random chance, which, in turn, would undermine the finding.

From the combined double high-low-portfolios, there is only a weakly significant alpha for the social dimension, 47 bps at the 10% level, where sustainable stocks appear to be dominant compared to less sustainable within the dispersion effect. The statistical power is, however, minimal, as both high and low dispersions are rather similar in magnitude and direction. For the other ESG components, alphas range between -15 and 22 bps but lack statistical significance. Taken together, European market shows inconclusive evidence of varying dispersion effect performance based on a firm's sustainability level.

For the U.S. market during the later 2016–2023 period, presented in Panel A of Table 8, we observe surprisingly similar trends in the governance portfolios as in the 2010–2015 period. The double high-low portfolio exhibits an alpha of -63 bps at the 5% level, continuing to suggest that stocks with low governance scores outperform those with high scores in the high-low dispersion setting. The effect persisting throughout the time frames is also evident in Panel C of Figure 3: the downward-sloping G high-low curve illustrates that the high dispersion performance is highly dependent on the level of governance a portfolio has.

Table 8: Two-Dimensional Sorting by Average ESG Rating and Rating Dispersion Level 2016–2023*. The table presents equally weighted FF5 model excess returns of portfolios sorted twice on ESG rating dispersion levels and ESG levels in the 2016–2023 period. The stocks are first sorted into quintiles based on ESG or its pillar rating dispersion, followed by a secondary sorting based on the average rating level of the respective component. Out of the nine portfolios (3x3), for brevity, we report only the results for the top (High) and bottom (Low) categories, as well as the long-short spread, each for both dimensions. Combined with the double High-Low-portfolio in the bottom right corners, this amounts to nine portfolios for each of the four pillar groups. All the numerical values are the FF5 model intercept coefficients and their t-statistics. Columns represent the ESG level, while rows indicate the level of rating dispersion based on the sample standard deviation of ratings from at least three providers out of MSCI, LSEG, Bloomberg, S&P Global. For instance, the cell labelled "High ESG, Low Disp." reports the excess returns of the highest ESG rating tercile within the low dispersion dimension tercile. T-statistics are adjusted with the Newey-West corrected standard errors, with values in parentheses. "*", "**", "***" denote statistical significance at levels 10%, 5%, and 1%, respectively. The interpretation of the table is explained in greater detail in the beginning of the section.*

Table 8 Panel A: U.S.

S&P 500	ESG			E			S			G		
	High	Low	High-Low	High	Low	High-Low	High	Low	High-Low	High	Low	High-Low
High Disp.	-0.02	-0.15	0.13	0.06	0.05	0.01	0.20	-0.25	**0.45****	-0.18	-0.01	-0.17
	(-0.18)	(-0.87)	(0.73)	(-0.37)	(-0.34)	(-0.03)	(1.18)	(-1.51)	(2.05)	(-1.14)	(-0.03)	(-0.85)
Low Disp.	-0.01	-0.36	0.35	-0.14	0.00	-0.13	0.16	-0.27	**0.43***	0.14	**-0.32****	**0.46****
	(-0.06)	(-1.56)	(1.47)	(-0.85)	(-0.02)	(-0.64)	(1.12)	(-1.30)	(1.75)	(0.88)	(-2.02)	(2.29)
High-Low Disp.	-0.01	0.21	-0.22	0.20	0.06	0.14	0.04	0.01	0.02	**-0.32****	0.32	**-0.63****
	(-0.09)	(1.04)	(-0.90)	(0.85)	(0.21)	(0.47)	(0.20)	(0.06)	(0.07)	(-2.06)	(1.61)	(-2.42)

Table 8 Panel B: Europe

STOXX 600	ESG			E			S			G		
	High	Low	High-Low	High	Low	High-Low	High	Low	High-Low	High	Low	High-Low
High Disp.	0.11	-0.21	**0.32***	0.03	-0.01	0.04	0.02	0.06	-0.04	0.12	**-0.38***	**0.50*****
	(0.57)	(-0.86)	(1.77)	(-0.15)	(0.04)	(-0.23)	(0.11)	(0.27)	(-0.26)	(0.56)	(-1.77)	(2.75)
Low Disp.	-0.02	-0.06	0.04	0.08	-0.14	0.22	0.00	-0.13	0.14	0.11	-0.15	0.26
	(-0.10)	(-0.27)	(0.17)	(0.46)	(-0.54)	(1.20)	(0.02)	(-0.66)	(0.78)	(0.44)	(-0.71)	(1.21)
High-Low Disp.	0.14	-0.15	0.28	-0.05	0.13	-0.19	0.01	0.20	-0.18	0.01	-0.23	0.24
	(0.90)	(-0.83)	(1.09)	(-0.14)	(0.30)	(-0.69)	(0.10)	(1.08)	(-0.69)	(0.08)	(-1.28)	(1.01)

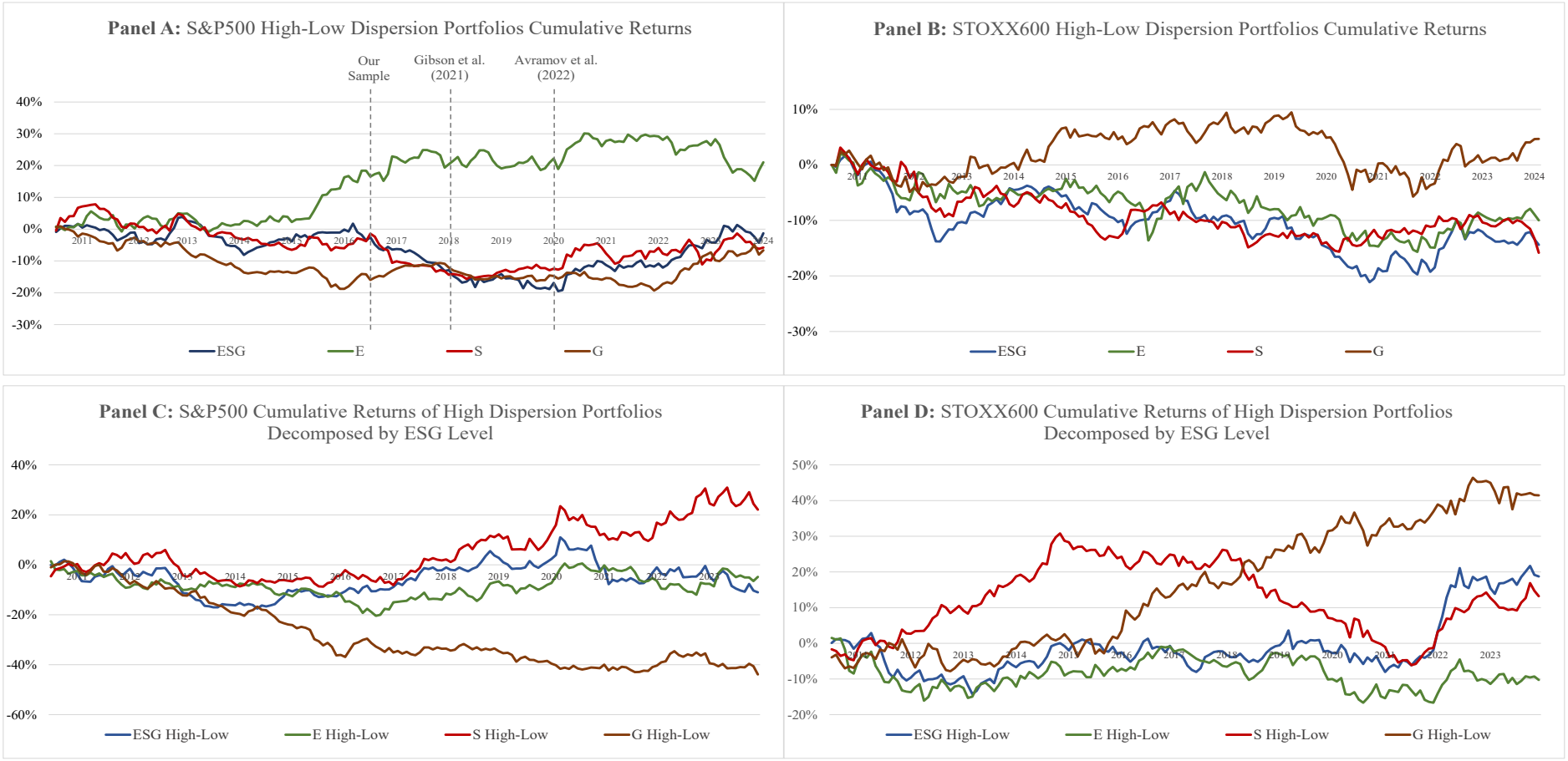
Since the coefficients of the bottom row portfolios in the governance pillar are within the same magnitude (0.32), but with opposite signs, the dispersion effect can cancel out without the second sorting. Indeed, the earlier Panel A of Table 6 shows no excess returns (10 bps) with stocks

of each governance tercile included. This, in and of itself, may indicate an interesting, albeit counter-intuitive, dynamic: rating disagreement only manifests in the case of high rating level. Essentially, disagreement in ratings only meaningfully disrupts pricing of stocks of high ratings. In practical terms, an unsustainable stock will be purchased whether the score is certain or uncertain. But with high sustainability, uncertainty leads to premium pricing. However, even if an interesting observation, we only find evidence for this relationship in the U.S. and for governance pillar sustainability.

Once more, practically none of the findings in the U.S. data extend to the European market during the later period, as shown in Panel B of Table 8. Alphas range from -19 bps to 28 bps for the double-sorted double high-low portfolios. However, the statistical power is insufficient to suggest any discernible effect in either direction. While the high dispersion in G favours sustainable stocks over less sustainable (50 bps at the 1% level), this effect does not carry over to the double high-low portfolio. Its direction is in fact opposite to the effect observed in both time frames in the U.S. market, where unsustainable stocks have an advantage. This is also visually evident from Panels C and D of Figure 3.

In general, the level of ESG and its components explain only a portion of the abnormal returns in the dispersion effect. We provide weak evidence that higher G score U.S. stocks are consistently explaining the negative dispersion effect in the governance dimension. Aside from this, however, we find no evidence of systematic divergence in the performance of sustainable versus unsustainable stocks across ESG dimensions in stocks with high or low disagreement.

Figure 3: Cumulative Returns for Long-Short Portfolios 2010–2023. *Figure 3 visually presents the cumulative returns for high-low dispersion portfolios based on the S&P 500 and STOXX 600 indices under two different settings. The top row shows the cumulative returns of the baseline portfolios (covered in 5.2 Portfolio Performance), while the bottom row presents the cumulative returns of high dispersion portfolios decomposed by the ESG level (covered in 5.3.1 Two-Dimensional Sort). For illustration, the first portfolio of "ESG high-low" in Panel C represents the high ESG score stock performance minus low ESG score performance within the high dispersion portfolio. In Panel A, the first vertical dashed line indicates the split point of our sample's return data. The subsequent two dashed lines mark the endpoints of the sample periods used in Gibson et al. (2021) and Avramov et al. (2022), respectively.* ***Note: the scales of y-axes returns (%) are not constant between panels, which is especially important for Panels B and C.***



5.3.2 Varying Methodologies

Our final tests act as robustness checks. To evaluate how portfolio results depend on the definition of dispersion, we apply three distinct methodological variations: non-standardized ratings, industry-adjusted dispersions, and pairwise rating dispersions. These serve as the main alternatives to our baseline approach, which relies on standardized dispersions. Furthermore, we examine the impact of varying return windows, where standardized dispersions are matched with lag periods of different length. For relevance and brevity, we focus specifically on the high-low dispersion portfolios and limit our review to the 2016–2023 period that has greater rating agency coverage. The tables for the 2010–2015 period are reported in the Appendix (Table A11). Altogether, the results from the varying methodologies regressions help us reconcile our results with findings of the prior research. We firstly compare the tables stand-alone, then contextualize them with prior studies.

For the different lag portfolios, we list only the coefficients of the one and 12-month lagged returns as well as a "January" portfolio, all reported in Table 9. The January column bases portfolios on the ESG rating dispersions of the December of the previous year. These constituents are then carried forward for the following year. Each portfolio is therefore only balanced once annually, in contrast with our other portfolios of monthly rebalancing. Therefore, a January strategy would incur the least trading costs. Gibson et al. (2021) report all their portfolio regressions with specifically the annually rebalanced January setting. However, as explained in Section 4.4., we find this strategy rather unimplementable in practice, as ESG ratings of even one, let alone multiple different providers, are rarely available the following month.

Beginning the review from the U.S. market, in Table 9, we find that our results do not change with small changes in methodology: overall, we find little to no evidence of meaningful abnormal performance. Yet, the values deviate between columns, which we mostly attribute to random chance: for instance, the governance dimension becomes seemingly significant in some of the columns, such as the 1-month and January portfolios, but then in others the effect disappears entirely. Given the inconclusive nature of the results and their insensitivity to methodological choices, our conclusion regarding the non-existent ESG dispersion effect remains unchanged for the U.S market.

Table 9: ESG Disagreement on Stock Returns 2016–2023 under varying methodologies. *The table reports equally weighted excess returns based on ESG rating disagreement across the Total (ESG), Environmental (E), Social (S), and Governance (G) dimensions. Column 3 ("Std") reports the baseline standardized dispersion results for comparability, extracted from Tables 5 & 6. Columns 4–6 contain different dispersion methods with the baseline 6-month return lag: "Non-Std" (non-standardized), "Industry" (industry-adjusted), and "Pairwise" (pairwise ratings). Columns 7–9 use the standardized dispersion method but with alternative return windows: "1-month" and "12-month" lags, and "January," which extends December dispersions over the following year. For January portfolios, the ESG rating dispersions of the preceding December are used to begin investing in January, with annual rebalancing. All results show CAPM and Fama-French five-factor "high-low" strategies, taking a long position on high and a short on low ESG disagreement stocks. "N" denotes portfolio observations. T-statistics are adjusted with the Newey-West corrected standard errors, with values in parentheses. "*", "**", "***" denote statistical significance at levels 10%, 5%, and 1%, respectively.*

Table 9 Panel A: U.S.

		Base	Dispersion Adjustments			Rolling Windows			
Dimension	Model	Std	Non-Std	Industry	Pairwise	1-Month	12-Month	January	N
ESG	CAPM	0.01	0.10	-0.05	0.18	0.02	0.05	0.01	96
		(0.06)	(0.64)	(-0.37)	(1.09)	(0.10)	(0.26)	(0.09)	
	FF5	-0.12	0.16	-0.11	0.05	-0.04	-0.06	-0.10	96
		(-0.67)	(1.04)	(-0.95)	(0.33)	(-0.29)	(-0.45)	(-0.72)	
E	CAPM	0.03	-0.01	-0.13	-0.03	0.04	-0.13	0.14	96
		(0.22)	(-0.03)	(-1.00)	(-0.22)	(0.27)	(-0.88)	(0.90)	
	FF5	0.10	0.05	-0.11	0.04	0.01	-0.09	0.14	96
		(0.60)	(0.25)	(-0.8)	(0.27)	(0.04)	(-0.65)	(0.96)	
S	CAPM	-0.07	0.07	-0.17	-0.08	-0.14	-0.15	-0.10	96
		(-0.38)	(0.35)	(-1.34)	(-0.48)	(-0.73)	(-0.82)	(-0.55)	
	FF5	-0.14	0.05	-0.18	-0.14	-0.14	-0.15	-0.16	96
		(-1.00)	(0.34)	(-1.5)	(-0.92)	(-1.02)	(-1.17)	(-1.28)	
G	CAPM	0.15	-0.09	**0.17***	0.15	**0.28****	-0.09	**0.24****	96
		(1.32)	(-0.61)	(1.71)	(1.32)	(2.53)	(-0.84)	(2.18)	
	FF5	0.12	-0.11	0.16	0.12	0.13	-0.09	0.17	96
		(1.14)	(-0.79)	(1.64)	(1.14)	(1.07)	(-0.82)	(1.45)	

For the European market in Panel B, the results appear even more susceptible to random noise. The governance component shows instead a negative CAPM alpha for the non-standardized portfolio (-21 bps at the 10% level), while no other methodological approach indicates outperformance in either direction. Some notable minor trends also observed in the U.S. persist here: for instance, 1-month and January portfolio figures are closely similar, which supports our critique of the methodology of Gibson et al. — the lag of 1 month is not realistic for implementation, a problem that also the January portfolio partially suffers from. Overall, much like in the U.S. market, we find no consistent dispersion effect across different methodological adjustments, and our conclusions remain unchanged. In other words, our results — or lack thereof — are robust for these variations in methodology.

Table 9 Panel B: Europe. *See table description above for explanation.*

		Base	Dispersion Adjustments			Rolling Windows			
Dimension	Model	Std	Non-Std	Industry	Pairwise	1-Month	12-Month	January	N
ESG	CAPM	-0.02	-0.05	0.05	0.00	0.12	-0.08	0.12	96
		(-0.15)	(-0.26)	(0.32)	(0.01)	(0.79)	(-0.46)	(0.78)	
	FF5	-0.01	-0.07	0.01	-0.04	0.13	-0.06	0.08	96
		(-0.13)	(-0.59)	(0.05)	(-0.37)	(1.10)	(-0.44)	(0.71)	
E	CAPM	0.00	-0.04	0.12	-0.05	0.06	-0.13	0.04	96
		(-0.02)	(-0.16)	(0.96)	(-0.26)	(0.36)	(-0.70)	(0.25)	
	FF5	0.09	-0.08	0.14	0.04	0.13	-0.03	0.16	96
		(0.57)	(-0.63)	(1.25)	(0.25)	(1.00)	(-0.22)	(1.14)	
S	CAPM	0.00	0.08	0.08	0.10	0.04	-0.13	-0.07	96
		(-0.02)	(0.46)	(0.67)	(0.85)	(0.37)	(-1.16)	(-0.61)	
	FF5	-0.01	0.00	0.02	0.04	0.01	-0.11	-0.10	96
		(-0.06)	(0.03)	(0.18)	(0.40)	(0.10)	(-0.98)	(-0.90)	
G	CAPM	-0.05	**-0.21***	-0.13	-0.04	-0.13	-0.18	-0.07	96
		(-0.34)	(-1.71)	(-0.98)	(-0.28)	(-0.94)	(-1.30)	(-0.47)	
	FF5	-0.01	-0.15	-0.06	0.01	-0.09	-0.07	-0.11	96
		(-0.09)	(-1.33)	(-0.63)	(0.07)	(-0.72)	(-0.65)	(-0.86)	

Even though the results of the varying methodology regression do not affect meaningfully our original results, they help us reconcile the findings to those of the two main papers on the issue, Gibson et al. (2021) and Avramov et al. (2022). The findings aligning the results are three-fold. These pertain to the main methodological differences in dispersion calculations: the industry-adjusted dispersion values, the pairwise sample standard deviation, and the difference between the two papers. To avoid overinterpretation, we present them more as conjectures, but important signals of underlying differences nonetheless. These findings are partially from the Panel A of Table A11, excluded to Appendix for brevity. Table A11 is the 2010–2015 period equivalent of Table 9 above.

Firstly, the industry-adjusted regression column helps align our methodology closer to that of Gibson et al. Indeed, following this change, all pillar score figures in Table A11 are within a similar range as in the original paper, apart from the total ESG dimension: Environmental pillar portfolio is the only one with statistical significance, in both samples at the 10% level, with intercept coefficients from 23 to 25 bps that strongly compare to the Gibson et al. figures of 21 to 25 bps. Social and governance portfolio risk-adjusted returns are effectively zero in both samples. Total ESG pillar portfolios — which we speculated in Gibson et al. case to be driven mostly by residual, spillover effects of the environmental pillar portfolios — are also positive (12 bps vs. 23 bps) in both samples, but do not reach significancy during our time frame[12].

[12] As presented in Figure 3, our former sample period is not exactly the same as that of Gibson et al., which could explain some of the remaining differences. Then again, neither are our rating agencies.

To help align our methodology to Avramov et al., next, we compare the base case dispersion results to pairwise dispersion results. For reasons explained in detail in Section 4.4, we do not have pairwise regressions in 2010–2015, and therefore we can only compare the pairwise regressions of the later period. This naturally creates a major mismatch between the samples, which complicates the comparison. However, Avramov's sample reaches nearly halfway to our later period as well, illustrated in Figure 3. Indeed, the comparison does signal of a methodological influence: Avramov et al. only study the portfolio performance of the total ESG dimension, with returns adjusted with CAPM, where they note a weakly significant alpha of 23 bps. In our sample the respective figure is a closely similar 18 bps.

However, the final coefficient in Avramov's total ESG dimension, 18 bps, is specifically from their full sample, which in their subsample analysis is shown to have a lower effect between 2011–2019. This contradicts the findings of Gibson et al., even though the two papers have a similar time period and rating agency coverage. We attribute this difference, once again, mostly to the choice of dispersion calculation methodology: Gibson et al. adjust for industries, while Avramov et al. calculate standard deviation pairwise. Comparison of the respective columns in Table 9 Panel A shows that the two adjacent coefficients differ strongly between the methodologies, even if Avramov et al. only report total ESG portfolio performance.

6 Discussion

In this section, we present a summary of our key findings and discuss their implications. We begin by outlining the results most relevant to the focus of our study. We also address the limitations and consider how these may influence the interpretation of our results. Finally, we propose potential avenues for future studies in the growing field of ESG rating dispersion research.

6.1 Result Summary & Implications

In the first part of our analysis, we confirm considerable ESG rating disagreement. In general, our findings are similar to those of earlier literature: First, the pairwise correlations we document, from 0.20–0.55, are in range with those of previous studies (e.g., Chatterji et al., 2016; Berg et al., 2022). We also find the general level of ESG rating disagreement, together with its pillar dimensions, to average at around 20 basis points both in the later and former sample period. We further note that significant disagreement persists across time, regions, and industries. Even though some variation of disagreement is noted in these dimensions, we fail to find systematic patterns or causes. Then again, a deeper investigation is beyond the scope of this study and therefore left for future research.

Whereas prior research has mostly explored these in the U.S. stock markets (see, e.g., Berg et al. 2022; Christensen et al. 2022), we find conformingly also in the European companies. There does not seem to be great differences between the two markets, even though slight variations, perhaps due to mere chance, are observed: Correlations are similar across the two markets, as are the distributions of rating grades. Disagreement characteristics are also closely similar, with slightly greater disagreement in Europe. We attribute this to the notion of Christensen et al. (2022) who effectively argue that with increasing transparency, more disagreement will follow — a pattern consistent with Europe having greater care for ESG matters. Lastly, differences between industries persist in both, even if there is slight variation which specific industries are the outliers in the rating disagreement of different dimensions. Without closer inspection on the regional differences in sector disagreement, we also attribute these differences to random chance. Altogether, disagreement, in a similar fashion, is substantial in both markets.

Next, we move to the summary and implications of the main section of our study. Collectively, we find that neither the S&P 500 nor the STOXX 600 regressions provide evidence of superior performance by the high dispersion portfolio over the low leg. The environmental dimension in the U.S. market shows only limited support for high disagreement stocks outperforming low disagreement ones between 2010 and 2015. In contrast, the governance dimension exhibits a more consistent but reverse pattern: low dispersion stocks significantly outperform their high dispersion counterparts across all models. The S&P 500 overall shows a similar outcome in the later 2016–2023 period as well, with no evidence of statistically or economically significant dispersion alpha. The earlier observations in the environmental and governance dimensions do not persist, which suggests that any dispersion-related abnormal returns have faded over time. These results are by and large robust for the differences in methodology choices.

In the European market, most high-low regressions during the 2010–2015 period do not yield excess returns in either direction. Although the governance dimension produces a weakly significant and positive result, the inconsistency across models makes the evidence strongly inconclusive. This lack of pattern extends to the 2016–2023 period, where none of the ESG dimensions exhibit significant high-low dispersion alpha under any model specification.

The methodological differences help explain the lack of uniformity between our findings and those of prior literature. With minor changes in especially dispersion calculations, we could better replicate the results more the two main papers, Gibson et al. (2021) and Avramov et al. (2022). These findings effectively undermine the statistical significance and robustness of their results. Their findings also contradict one another: in the portfolio sorts of Gibson et al., they find excess returns in both total ESG and environmental dimensions between 2010 and 2017. Meanwhile Avramov et al., in their full sample, use the sole total ESG score and identify an alpha significant at the 10 % level, only for the effect to fade away after 2011. Therefore, the findings of these two papers do not strongly support one another — or their proposed theories behind it, for that matter. As such, we believe the effect of ESG disagreement is, to an extent, overstated and overinterpreted. Our results, which focus from 2010 onward, do not strictly contradict the pillar findings of Gibson et al., but we also fail to find continuous evidence in support of them either.

Of course, since the rating providers across the study samples are not identical, especially for the first period, any dispersion effect may be sensitive to the mix of rating agencies included. Still, the four providers we use in our later sample — three of which are directly used by both comparable papers[13] — are likely the most prominent actors in the markets (Wong et al., 2019), whose ratings are therefore most used, and who hence proxy the dispersion in the market best.

Even if the shared period of 2010–2015 would have had an effect of ESG rating dispersion, our sample indicates for this to have later disappeared. One possible theoretical explanation for the shift, as suggested by Avramov et al. (2022), lies in the increasing presence of green preferences among investors. These preferences imply that non-monetary benefits associated with ESG investing may reduce the need for a financial risk premium. If such benefits have grown in relevance over time, they could help explain why alpha associated with high ESG disagreement seems to disappear in later periods. Essentially, with growing relevance of ESG matters in investments decisions, even noisy ESG ratings become acceptable sources of information, regardless of any inherent ESG uncertainty or risk.

Our slight detour into two-dimensional portfolio sorting reveals limited and inconsistent evidence that ESG rating dispersion affects sustainable and unsustainable stocks differently. Some isolated results, such as strong returns low dispersion environmental stocks in Europe and highly rated stocks explaining the dispersion effect in the European social dimension, do appear, but they lack robustness and consistency. Overall, ESG level does not systematically moderate the effect of rating disagreement. The dispersion premium — if it exists — seems confined to specific settings, for instance to less sustainable governance U.S. stocks, which explain the dispersion effect persistently in both periods. Nevertheless, the double-sorting presents an interesting new avenue for academia, one we hope will be explored further.

Altogether, if the ESG rating dispersion effect was consistent, we would expect each ESG component to influence stock returns based on its level of dispersion. For dispersion to be considered as systematic risk to be priced in the market, ESG uncertainty should be clearly associated with higher expected returns. This should not limit to the total ESG pillar but instead be visible in all three of its subdimensions as well, provided they hold weight for investment decisions. However, our results suggest that the effect is at best inconsistent and temporary, potentially

[13] Major consolidation efforts make a direct comparison difficult; see detailed description in Section 3.2.

arbitrarily mean-reversing. Of course, the environmental dimension, and by extension the total ESG pillar, could be of greater interest to the average investor, but it is still difficult to argue that social and governance pillars would have no effect — much less a contrary relationship.

The lack of a measurable dispersion effect across ESG dimensions also calls into question the reliability of the ESG ratings themselves. The varying rating methodologies, scales, distributions, and other rating inconsistencies potentially blur any underlying relationship between ESG dispersion and stock returns. Therefore, the results we observe — or fail to observe — might not have accurately captured true ESG rating disagreement in the first place. The ratings could capture noise just as much as they reflect genuine ESG performance. This naturally has its own implications for ESG research that bases the studies in solely a single provider's ESG ratings — would an observed sustainability effect ever have appeared, had the authors had access to a different source of ESG information?

6.2 Limitations

6.2.1 Rating Data & Characteristics

Although the limited number of data vendors for the 2010–2015 period presents a clear limitation, we believe that even only the two existing data sources do both provide sufficient coverage and proxy disagreement well enough to gauge its effects. Fundamentally, the disagreement in ESG ratings extends beyond a simple difference of two scores; it reveals internal contradictions within each agency's analysis process**.** In other words, while there are only two ratings, the underlying individual analyst dispersion is much greater. Therefore, we believe the extended sample remains valuable and worthy of study. That still is not to say that the earlier section would not benefit from better coverage, however.

Another widely recognized limitation in ESG research is the inconsistency in rating methodologies among rating agencies, as discussed also in the literature review (see., e.g., Chatterji et al., 2016; Berg et al., 2022). The absence of standardized measures may result in agencies arriving at vastly different conclusions regarding a company's ESG performance. While lack of standardization is essentially at heart of what we are investigating, more arbitrary grading

naturally obscures the effect of specifically rating disagreement and uncertainty, or compensation for market risk.

Another issue is the potential change in rating methods within our study's time frame. Given that global expectations for ESG compliance are constantly increasing and rating frameworks developing, it is likely that the rating agencies have tightened or otherwise revised their methodologies. We try to counteract this by dividing the time frame into two parts, where the cutoff coincides with major sustainability landmarks such as the signing of Paris Agreement and establishment of the Sustainable Development Goals. Nevertheless, the ratings are far from stable, especially during major consolidation efforts in the rating industry. Changes of rating methodologies can, at the very least, cloud what part of ESG dispersion is due to ESG rating noise, and what part of it is for actual ESG rating disagreement. These would therefore effectively worsen ESG rating dispersion's ability to proxy disagreement and by extension, dilute any stock market effects.

6.2.2 Data Coverage

The study is naturally limited to the time period choice of 2010 to 2023. While including earlier rating data increases the number of observations and therefore could improve the reliability of the results, the percentage of rated companies substantially decreases the farther we move away from the present. Hence, we face a trade-off between time and data coverage. The effective sample size is further limited by the infrequent rating updates, as majority of the ratings are revised only on an annual basis.

Naturally, the study is also narrowed down to only include the stock markets of the U.S. and Europe. This could make the results not applicable to other geographical locations. The mere comparison between the two may not be entirely valid, as evidenced by the differing findings. Indeed, indices proxying the markets are not perfectly comparable. Different markets may price ESG and its disagreement completely different, which means thoughtless extrapolation may not be justified. Within our sample, we address this by comparing the S&P 500 and STOXX 600, while broader geographical expansions are left for future research. In general, our thesis includes other additional limitations, of which we aimed to address the main few through robustness checks; even then, these checks are by no means fully definitive.

6.3 Further research

Nearly each and every paper calls for standardization of rating methodologies, yet very few investigate its effects. We hope later studies focus on the convergence of ratings, what has and would make them converge. In a similar vein, we are interested in knowing why correlations or disagreements between agencies have not decreased in the past two decades, despite major advances in sustainability. Are Christensen et al. (2022) in the right claiming that more disclosure exacerbates disagreement, or could the opposing idea of Kimbrough et al. (2024) carry more weight in the future?

For the lack of research agreement of the effects, future research should investigate theoretical explanations why there does not exist any return premium for ESG rating disagreement. We theorize that it could be for instance because of two counterbalancing forces of heterogenous beliefs in the cross-section, but do not further explore it empirically. Yet, as sustainability matters are already of major interest to market participants, we find it hard to believe that ESG disagreement would not have a stronger effect. It is especially difficult to accept that newer samples find little to no evidence confirming the past results. This instead indicates that the effects of disagreement decrease whilst the importance of ESG is growing.

On the methodological side, studies could test different ways to standardize ESG scores and better track how disagreement changes over time. Moreover, instead of merely focusing on market returns, researchers could better examine how ESG disagreement relates to volatility, trading volume, stock price crash risk and crash impact on price. Lastly, while we include factor loadings of ESG disagreement regressions in the Appendix, their analysis is beyond our scope, and thus, we hope academics will continue where we left off.

Lastly, more attention should be given to ESG disagreement effect within different ESG rating levels to confirm, for instance, whether environmental rating disagreement only has stock market implications within “green” stocks. We only scratch the surface with our tests but recognize interesting early signs nonetheless. Likewise, research could also explore whether behavioural factors play a role, much like investor sentiments for the effect of company fundamentals. Certain green sentiment indices — such as the media climate change concerns, “MCCC”, of Ardia et al. (2023) — could be of interest for specifically ESG rating disagreement.

7 Conclusion

Our study contributes to the lack of agreement on the research of ESG rating disagreement. Despite the seemingly conforming results of the prior research, at least on surface level, we fail to replicate their results. With "surface level" we imply that the major previous papers on the matter may have overstated or overinterpreted the effect. Our findings indicate that ESG rating dispersion does not consistently affect stock returns in the U.S. and European stock markets, which challenges the notion that disagreement in ESG ratings inherently leads to higher expected returns. Our evidence seems to best support the tangential proposition of Harvey et al. (2016): "most claimed research findings in financial economics are likely false."

We attribute the lack of excess returns of dispersion portfolios on three different financial theories. The main two pertain to heterogenous beliefs in the stock market that we then extend to ESG dimension. First posits that stock market disagreement causes short-term overvaluation and subsequently lower returns. Some authors argue instead the contrary: disagreement causes uncertainty, translating to greater risk, which should then be compensated for in higher returns. We believe, irrespective of contradicting implications, the two can coexist and interact. In the cross-section, the two theories cloud the effects of one another. General stock market noise would likely hide any remainder effect. With ESG ratings the noise is naturally amplified for the lag between updated ratings and portfolio returns. The final theory, which helps reconcile the mixed findings of our two time periods, is proposed by Avramov et al. (2022): green market sentiments dilute the effects of uncertainty resulting from dispersion. Simplified, when sustainable assets are increasingly sought for, even noisy ratings are acceptable.

Nevertheless, the ultimate reasons for the lack of a "dispersion premium" remain uncertain. But for investors, the implications are rather clear: relying on ESG rating dispersion as a standalone signal for stock selection would likely not provide consistent outperformance. Even if premium on dispersion portfolios can be identified with certain rating agency compositions, ESG dimensions, market regions or time frames, it is likely difficult to replicate and transform into a viable investment strategy ex post. Then again, the dispersion effect is certainly neither a dispersion discount. Whereas high ESG levels can detriment financial performance, our ESG dispersion portfolios are on par with average investment strategies. A sustainable investor could therefore benefit from high ESG dispersion, just in a different way than previously theorized. This idea, however, is left for future academic endeavours.

8 References

Albuquerque, R., Koskinen, Y., & Zhang, C. (2019). Corporate social responsibility and firm risk: Theory and empirical evidence. *Management science*, 65(10), 4451-4469.¨

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

Anderson, E. W., Ghysels, E., & Juergens, J. L. (2005). Do heterogeneous beliefs matter for asset pricing?. *The Review of Financial Studies*, *18*(3), 875-924.

Ardia, D., Bluteau, K., Boudt, K., & Inghelbrecht, K. (2023). Climate change concerns and the performance of green vs. brown stocks. *Management Science*, *69*(12), 7607-7632.

Atmaz, A., & Basak, S. (2018). Belief dispersion in the stock market. *The Journal of Finance*, 73(3), 1225-1279.

Atz, U., Van Holt, T., Liu, Z. Z., & Bruno, C. C. (2023). Does sustainability generate better financial performance? review, meta-analysis, and propositions. *Journal of Sustainable Finance & Investment*, *13*(1), 802-825.

Avramov, D., Cheng, S., Lioui, A., & Tarelli, A. (2022). Sustainable investing with ESG rating uncertainty. *Journal of financial economics*, 145(2), 642-664.

Baker, M., Bergstresser, D., Serafeim, G., & Wurgler, J. (2022). The pricing and ownership of US green bonds. *Annual review of financial economics*, *14*(1), 415-437.

Barberis, N., & Thaler, R. (2003). A survey of behavioral finance. *Handbook of the Economics of Finance*, *1*, 1053-1128.

Berg, F., Koelbel, J. F., & Rigobon, R. (2022). Aggregate confusion: The divergence of ESG ratings. *Review of Finance,* 26(6), 1315-1344.

Billio, M., Costola, M., Hristova, I., Latino, C., & Pelizzon, L. (2021). Inside the ESG ratings:(Dis) agreement and performance. *Corporate Social Responsibility and Environmental Management*, 28(5), 1426-1445.

Bloomberg L.P. (2025). ESG Disclosure Scores [data set]. *Bloomberg Terminal*. Accessed: 1/2025

Bloomberg L.P. (2025). S&P Global ESG Scores [data set]. *Bloomberg Terminal*. Accessed: 1/2025

Cantor, R., & Packer, F. (1995). The credit rating industry. The Journal of Fixed Income, 5(3), 10-34.

Capizzi, V., Gioia, E., Giudici, G., & Tenca, F. (2021). The divergence of ESG ratings: An analysis of Italian listed companies. *Journal of Financial Management, Markets and Institutions,* 9(02), 2150006.

Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of finance*, *52*(1), 57-82.

Chatterji, A. K., Durand, R., Levine, D. I., & Touboul, S. (2016). Do ratings of firms converge? Implications for managers, investors and strategy researchers. *Strategic management journal*, 37(8), 1597-1614.

Chatterji, A. K., Levine, D. I., & Toffel, M. W. (2009). How well do social ratings actually measure corporate social responsibility?. *Journal of Economics & Management Strategy*, 18(1), 125-169.

Chen, S., Song, Y., & Gao, P. (2023). Environmental, social, and governance (ESG) performance and financial outcomes: Analyzing the impact of ESG on financial performance. *Journal of environmental management*, 345, 118829.

Cheng, B., Ioannou, I., & Serafeim, G. (2014). Corporate social responsibility and access to finance. *Strategic Management Journal*, 35(1), 1-23.

Christensen, D. M., Serafeim, G., & Sikochi, A. (2022). Why is corporate virtue in the eye of the beholder? The case of ESG ratings. *The Accounting Review*, 97(1), 147-175.

Cochran, P. L., & Wood, R. A. (1984). Corporate social responsibility and financial performance. *Academy of management Journal*, 27(1), 42-56.

Cornell, B. (2021). ESG preferences, risk and return. *European Financial Management,* 27(1), 12-19.
Datastream International [data set]. *LSEG Workspace*. Accessed: 1/2025

Delmas, M., & Blass, V. D. (2010). Measuring corporate environmental performance: the trade-offs of sustainability ratings. *Business Strategy and the Environment*, *19*(4), 245-260.

Diether, K. B., Malloy, C. J., & Scherbina, A. (2002). Differences of opinion and the cross section of stock returns. *The journal of finance*, *57*(5), 2113-2141.

Dimson, E., Marsh, P., & Staunton, M. (2020). Divergent ESG ratings. *Portfolio management research*

Dong, M., Li, M., Wang, H., & Pang, Y. (2025). ESG disagreement and stock price crash risk: evidence from China. *Asia-Pacific Financial Markets*, 32(1), 267-299.

Engelberg, J. E., Reed, A. V., & Ringgenberg, M. C. (2018). Short-selling risk. *The Journal of Finance*, 73(2), 755-786.

Eugene, F., & French, K. (1992). The cross-section of expected stock returns. *Journal of finance*, *47*(2), 427-465.

Fama, E. F. (1970). Efficient capital markets. *Journal of finance*, *25*(2), 383-417.

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, *33*(1), 3-56.

Fama, E. F., & French, K. R. (2007). Disagreement, tastes, and asset prices. *Journal of financial economics*, *83*(3), 667-689.

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

Fama, E. F., & MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. *Journal of political economy*, *81*(3), 607-636.

Feng, J., Goodell, J. W., & Shen, D. (2022). ESG rating and stock price crash risk: Evidence from China. Finance Research Letters, 46, 102476.

Friede, G., Busch, T., & Bassen, A. (2015). ESG and financial performance: aggregated evidence from more than 2000 empirical studies. Journal of sustainable finance & investment, 5(4), 210-233.

Galant, A., & Cadez, S. (2017). Corporate social responsibility and financial performance relationship: A review of measurement approaches. *Economic research-Ekonomska istraživanja*, *30*(1), 676-693.

Gibson Brandon, R., Krueger, P., & Schmidt, P. S. (2021). ESG rating disagreement and stock returns. *Financial analysts journal*, 77(4), 104-127.

Giglio, S., Maggiori, M., Stroebel, J., Tan, Z., Utkus, S., & Xu, X. (2025). Four facts about ESG beliefs and investor portfolios. *Journal of Financial Economics*, 164, 103984.

Griffin, J. J., & Mahon, J. F. (1997). The corporate social performance and corporate financial performance debate: Twenty-five years of incomparable research. *Business & society*, *36*(1), 5-31.

Haji, A. A., Coram, P., & Troshani, I. (2022). Consequences of CSR reporting regulations worldwide: a review and research agenda. *Accounting, Auditing & Accountability Journal*, 36(1), 177-208.

Hartzmark, S. M., & Sussman, A. B. (2019). Do investors value sustainability? A natural experiment examining ranking and fund flows. *Journal of Finance,* 74(6), 2789-2837.

Harvey, C. R., Liu, Y., & Zhu, H. (2016). … and the cross-section of expected returns. *Review of Financial Studies*, 29(1), 5-68

Heinkel, R., Kraus, A., & Zechner, J. (2001). The effect of green investment on corporate behavior. *Journal of Financial and Quantitative Analysis*, 36(4), 431-449.

Henisz, W., Koller, T., & Nuttall, R. (2019). Five ways that ESG creates value*. McKinsey Quarterly*, 4, 1-12.

Hirai, A., & Brady, A. (2021, July 28). Managing ESG data and rating risk. *Harvard Law School Forum on Corporate Governance.* https://corpgov.law.harvard.edu/2021/07/28/managing-esg-data-and-rating-risk/. Accessed: 17.3.2025

Hong, H., & Kacperczyk, M. (2009). The price of sin: The effects of social norms on markets. *Journal of financial economics*, *93*(1), 15-36.

Ilinitch, A. Y., Soderstrom, N. S., & Thomas, T. E. (1998). Measuring corporate environmental performance. *Journal of accounting and public policy*, *17*(4-5), 383-408.

Jensen, M. C. (1968). The performance of mutual funds in the period 1945-1964. *The Journal of finance,* 23(2), 389-416.

Jones, C. M., & Lamont, O. A. (2002). Short-sale constraints and stock returns. *Journal of Financial Economics*, *66*(2-3), 207-239.

Khan, M. A. (2022). ESG disclosure and firm performance: A bibliometric and meta analysis. *Research in International Business and Finance*, 61, 101668.

Eugene, F., & French, K. (1992). The cross-section of expected stock returns. *Journal of Finance*, 47(2), 427-465.

Khan, M., Serafeim, G., & Yoon, A. (2016). Corporate sustainability: First evidence on materiality. *The accounting review,* 91(6), 1697-1724.'

Kim, Y., Li, H., & Li, S. (2014). Corporate social responsibility and stock price crash risk. *Journal of Banking & Finance,* 43, 1-13.

Kimbrough, M. D., Wang, X., Wei, S., & Zhang, J. (2024). Does voluntary ESG reporting resolve disagreement among ESG rating agencies?. *European Accounting Review,* 33(1), 15-47.

Knight, F. H. (1921). *Risk, uncertainty and profit* (Vol. 31). Houghton Mifflin.

Lim, K. P., & Brooks, R. (2011). The evolution of stock market efficiency over time: A survey of the empirical literature. *Journal of economic surveys*, *25*(1), 69-108.

Lins, K. V., Servaes, H., & Tamayo, A. (2017). Social capital, trust, and firm performance: The value of corporate social responsibility during the financial crisis. *Journal of Finance*, *72*(4), 1785-1824.

Lintner, J. (1965). Security prices, risk, and maximal gains from diversification. *Journal of Finance*, 20(4), 587-615.

Liu, X., Yang, Q., Wei, K., & Dai, P. F. (2024). ESG rating disagreement and idiosyncratic return volatility: Evidence from China. *Research in International Business and Finance*, 70, 102368.

London Stock Exchange Group (2024) LSEG ESG Scores Methodology. Retrieved from: https://www.lseg.com/content/dam/data-analytics/en_us/documents/methodology/lseg-esg-scores-methodology.pdf. Accessed: 29.4.2025

Luo, D., Yan, J., & Yan, Q. (2023). The duality of ESG: Impact of ratings and disagreement on stock crash risk in China. *Finance Research Letters*, 58, 104479.

Mackey, A., Mackey, T. B., & Barney, J. B. (2007). Corporate social responsibility and firm performance: Investor preferences and corporate strategies. *Academy of management review*, *32*(3), 817-835.

Martin, I. W., & Nagel, S. (2022). Market efficiency in the age of big data. *Journal of financial economics,* 145(1), 154-177.

McWilliams, A., & Siegel, D. (2000). Corporate social responsibility and financial performance: correlation or misspecification?. *Strategic management journal*, 21(5), 603-609.

Miller, E. M. (1977). Risk, uncertainty, and divergence of opinion. *The Journal of finance*, *32*(4), 1151-1168.

MSCI (2024) MSCI ESG Ratings Methodology. Retrieved from: https://www.msci.com/documents/1296102/34424357/MSCI+ESG+Ratings+Methodology.pdf. Accessed: 29.4.2025

MSCI Inc. ESG ratings [Data set]. From https://www.msci.com/our-solutions/esg-investing/esg-ratings. Accessed: 1/2025

Nelling, E., & Webb, E. (2009). Corporate social responsibility and financial performance: The "virtuous circle" revisited. *Review of Quantitative finance and accounting*, 32, 197-209.

Newey, W. K., & West, K. D. (1986). A simple, positive semi-definite, heteroskedasticity and autocorrelation-consistent covariance matrix.

Nollet, J., Filis, G., & Mitrokostas, E. (2016). Corporate social responsibility and financial performance: A non-linear and disaggregated approach. *Economic Modelling*, *52*, 400-407.

Orlitzky, M., Schmidt, F. L., & Rynes, S. L. (2003). Corporate social and financial performance: A meta-analysis. *Organization studies*, *24*(3), 403-441.

Pástor, Ľ., Stambaugh, R. F., & Taylor, L. A. (2021). Sustainable investing in equilibrium. Journal of financial economics, 142(2), 550-571.

Pástor, Ľ., Stambaugh, R. F., & Taylor, L. A. (2022). Dissecting green returns. *Journal of financial economics*, *146*(2), 403-424.

Pedersen, L. H., Fitzgibbons, S., & Pomorski, L. (2021). Responsible investing: The ESG-efficient frontier. *Journal of financial economics*, *142*(2), 572-597.

Preston, L. E. (Ed.). (1978). Research in corporate social performance and policy. Greenwich, CT: JAI Press.

Principles for Responsible Investment. (2024). PRI Annual Report 2024. Principles for Responsible Investment. Retrieved from: https://www.unpri.org/download?ac=21536

Rau, P. R., & Yu, T. (2024). A survey on ESG: investors, institutions and firms. *China Finance Review International*, 14(1), 3-33.

Riedl, A., & Smeets, P. (2017). Why do investors hold socially responsible mutual funds?. *the Journal of Finance*, *72*(6), 2505-2550.

S&P Global (2024) S&P Global ESG Scores Methodology. Retrieved from: https://portal.s1.spglobal.com/survey/documents/spglobal_esg_scores_methodology.pdf. Accessed: 11.3.2025

Sadka, R., & Scherbina, A. (2007). Analyst disagreement, mispricing, and liquidity. *The Journal of Finance,* 62(5), 2367-2403.

Semenova, N., & Hassel, L. G. (2015). On the validity of environmental performance metrics. *Journal of business ethics*, *132*, 249-258.

Serafeim, G., & Yoon, A. (2023). Stock price reactions to ESG news: The role of ESG ratings and disagreement. *Review of Accounting Studies*, 28(3), 1500-1530.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, *19*(3), 425-442.

STOXX. STOXX Europe 600 Index. from https://www.stoxx.com/index-details?symbol=SXXP. Accessed: 28.3.2025

The Good Lobby (2024) TGL Scorecard – Bloomberg. Retrieved from: https://www.thegoodlobby.eu/wp-content/uploads/2024/12/TGL-Scorecard-Bloomberg.pdf. Accessed: 29.4.2025

Wang, J., Wang, S., Dong, M., & Wang, H. (2024). ESG rating disagreement and stock returns: Evidence from China. *International Review of Financial Analysis*, 91, 103043.

Wong, C., A. Brackley, E. Petroy. 2019. Rate the Raters 2019: Expert Views on ESG Ratings. Sustainability. https://www.erm.com/globalassets/sustainability.com/thinking/pdfs/sa-ratetheraters-2019-1.pdf. Accessed: 28.1.2025

Yen, G., & Lee, C. F. (2008). Efficient market hypothesis (EMH): past, present and future. *Review of Pacific Basin Financial Markets and Policies*, *11*(02), 305-329.

Zeng, Q., Xu, Y., Hao, M., & Gao, M. (2025). ESG rating disagreement, volatility, and stock returns. *Finance Research Letters,* 72, 106602.

9 Appendix

9.1 Factor Loadings

Table A10: Factor Loadings for ESG Disagreement Portfolios, 2016–2023. *The table reports factor loadings based on the Fama-French five-factor model for ESG rating disagreement portfolios across the Total ESG, Environmental (E), Social (S), and Governance (G) dimensions. Columns 3–8 display the coefficients for the abnormal return ("Intercept"), excess market return ("MKT-RF"), Small Minus Big ("SMB"), High Minus Low ("HML"), Robust Minus Weak ("RMW"), and Conservative Minus Aggressive ("CMA") factors. The "Adj-R²" column reports the adjusted R-squared value of the regression. The portfolio "High" represents the highest disagreement, while "Low" the smallest. The "High-Low" row denotes a strategy with a long position on high ESG disagreement stocks and a short position on low ESG disagreement stocks. T-statistics are adjusted with the Newey-West corrected standard errors, with values in parentheses. "*", "**", "***" denote statistical significance at levels 10%, 5%, and 1%, respectively. Note: this table is an extension of Table 6.*

Table A10 Panel A: U.S.

Dimension	Portfolio	INTERCEPT	MKT-RF	SMB	HML	RMW	CMA	Adj-R^2
ESG	High	-0.17	**1.03*** **	**0.19*** **	**0.11** **	**0.21*** **	**0.17** **	0.94
		(-1.22)	(31.32)	(3.07)	(2.18)	(2.85)	(2.17)	
	Low	-0.05	**1.03*** **	**0.14** **	**0.27*** **	0.04	-0.02	0.95
		(-0.36)	(32.02)	(2.37)	(5.39)	(0.56)	(-0.28)	
	High-Low	-0.12	0.00	0.05	**-0.16*** **	**0.17***	**0.19** **	0.05
		(-0.67)	(0.06)	(0.65)	(-2.63)	(1.96)	(2.08)	
E	High	0.03	**1.02*** **	**0.15*** **	**0.24*** **	0.07	-0.03	0.96
		(0.21)	(36.53)	(2.87)	(5.43)	(1.04)	(-0.39)	
	Low	-0.07	**0.97*** **	0.10	**0.21*** **	**0.17** **	-0.01	0.93
		(-0.42)	(28.53)	(1.48)	(3.89)	(2.20)	(-0.11)	
	High-Low	0.10	0.05	0.06	0.03	-0.10	-0.02	0.06
		(0.60)	(1.43)	(0.88)	(0.57)	(-1.36)	(-0.21)	
S	High	-0.12	**1.01*** **	**0.14** **	**0.12** **	**0.19*** **	0.03	0.94
		(-0.94)	(31.12)	(2.32)	(2.47)	(2.64)	(0.44)	
	Low	0.02	**1.00*** **	**0.21*** **	**0.32*** **	0.05	0.03	0.96
		(0.16)	(35.12)	(3.82)	(7.21)	(0.81)	(0.52)	
	High-Low	-0.14	0.01	-0.06	**-0.19*** **	**0.14** **	0.00	0.35
		(-1.00)	(0.37)	(-1.12)	(-4.19)	(2.10)	(-0.01)	
G	High	0.06	**0.98*** **	**0.11** **	**0.23*** **	**0.10***	0.09	0.96
		(0.51)	(39.88)	(2.27)	(5.99)	(1.82)	(1.57)	
	Low	-0.06	**1.01*** **	**0.14** **	**0.19*** **	**0.12***	0.01	0.95
		(-0.54)	(34.53)	(2.51)	(4.13)	(1.86)	(0.15)	
	High-Low	0.12	-0.03	-0.03	0.04	-0.02	0.08	0.07
		(1.14)	(-1.28)	(-0.68)	(0.99)	(-0.38)	(1.30)	

Table A10 Panel B: Europe

Dimension	Portfolio	INTERCEPT	MKT-RF	SMB	HML	RMW	CMA	Adj-R^2
ESG	High	-0.03	**0.78*****	0.07	**0.51*****	0.28	-0.22	0.87
		(-0.14)	(18.16)	(0.58)	(3.89)	(1.57)	(-0.99)	
	Low	-0.01	**0.80*****	0.18	**0.27****	**0.33***	-0.27	0.86
		(-0.06)	(18.21)	(1.40)	(1.97)	(1.82)	(-1.23)	
	High-Low	-0.01	-0.02	-0.11	**0.25*****	-0.05	0.06	0.51
		(-0.13)	(-0.91)	(-1.58)	(3.47)	(-0.54)	(0.48)	
E	High	-0.01	**0.80*****	0.22	**0.43*****	0.09	-0.35	0.86
		(-0.03)	(17.16)	(1.59)	(2.99)	(0.48)	(-1.48)	
	Low	-0.09	**0.81*****	0.19	**0.37****	**0.64*****	-0.28	0.83
		(-0.40)	(15.85)	(1.26)	(2.40)	(3.03)	(-1.10)	
	High-Low	0.09	-0.01	0.03	0.05	**-0.55*****	-0.07	0.34
		(0.57)	(-0.23)	(0.30)	(0.53)	(-4.00)	(-0.40)	
S	High	0.06	**0.77*****	0.15	**0.43*****	**0.38****	-0.28	0.87
		(0.29)	(18.06)	(1.16)	(3.29)	(2.13)	(-1.28)	
	Low	0.06	**0.81*****	**0.25***	**0.44*****	**0.37****	-0.26	0.87
		(0.31)	(17.93)	(1.88)	(3.21)	(1.99)	(-1.16)	
	High-Low	-0.01	-0.03	-0.10	-0.01	0.01	-0.01	0.04
		(-0.06)	(-1.34)	(-1.37)	(-0.14)	(0.07)	(-0.11)	
G	High	-0.07	**0.78*****	0.19	**0.51*****	**0.43****	-0.16	0.87
		(-0.31)	(18.46)	(1.53)	(3.96)	(2.45)	(-0.74)	
	Low	-0.05	**0.78*****	0.21	**0.41*****	**0.46****	**-0.46****	0.87
		(-0.26)	(17.62)	(1.61)	(2.98)	(2.49)	(-2.03)	
	High-Low	-0.01	-0.01	-0.02	0.11	-0.03	**0.30****	0.37
		(-0.09)	(-0.21)	(-0.26)	(1.29)	(-0.28)	(2.22)	

9.2 Varying methodologies 2010–2015

Table A11: ESG Disagreement on Stock Returns 2010–2015 under varying methodologies. *The table reports equally weighted excess returns based on ESG rating disagreement across the Total (ESG), Environmental (E), Social (S), and Governance (G) dimensions. Column 3 ("Std") reports the baseline standardized dispersion results for comparability. Columns 4 and 5 contain different dispersion methods with the baseline 6-month return lag: "Non-Std" (non-standardized), "Industry" (industry-adjusted). Previously reported "Pairwise" column is excluded, as for 2010–2015, the portfolio constituents match the base case; see Section 4.4 for detailed explanation. Columns 7–9 use the standardized dispersion method but with alternative return windows: "1-month" and "12-month" lags, and "January," which extends December dispersions over the following year. All results show the intercept coefficients of the CAPM or FF5 "high-low" strategies, taking a long position on high and short on low ESG disagreement stocks. "N" denotes the portfolio monthly time-series observations. T-statistics are adjusted with the Newey-West corrected standard errors, with values in parentheses. "*", "**", "***" denote statistical significance at levels 10%, 5%, and 1%, respectively.*

Table A11 Panel A: U.S.

		Base	Dispersion Adjustments		Rolling Windows			
Dimension	Portfolio	Std	Non-Std	Industry	1-Month	12-Month	January	N
ESG	CAPM	0.02	0.12	0.12	0.13	-0.15	-0.03	72
		(0.11)	(0.79)	(1.07)	(1.00)	(-1.14)	(-0.21)	
	FF5	-0.06	-0.03	0.06	0.07	-0.18	-0.08	72
		(-0.43)	(-0.25)	(0.5)	(0.53)	(-1.35)	(-0.46)	
E	CAPM	0.22	0.11	**0.23***	0.02	-0.19	0.02	72
		(1.38)	(0.68)	(1.78)	(0.15)	(-1.10)	(0.11)	
	FF5	**0.30***	0.20	**0.25***	0.04	-0.11	0.02	72
		(1.92)	(1.22)	(1.89)	(0.36)	(-0.61)	(0.13)	
S	CAPM	-0.14	0.16	0.01	-0.03	**-0.26***	-0.04	72
		(-0.91)	(0.93)	(0.05)	(-0.20)	(-1.89)	(-0.26)	
	FF5	-0.21	0.00	-0.04	-0.08	**-0.29****	-0.09	72
		(-1.33)	(0.01)	(-0.29)	(-0.57)	(-2.12)	(-0.65)	
G	CAPM	**-0.26****	-0.19	-0.06	-0.01	-0.19	-0.13	72
		(-2.16)	(-1.46)	(-0.54)	(-0.12)	(-1.27)	(-1.07)	
	FF5	**-0.33*****	-0.19	-0.12	-0.03	**-0.24***	-0.15	72
		(-2.74)	(-1.55)	(-1.15)	(-0.26)	(-1.68)	(-1.28)	

Table A11 Panel B: Europe

		Base	Dispersion Adjustments		Rolling Windows			
Dimension	Portfolio	Std	Non-Std	Industry	1-Month	12-Month	January	N
ESG	CAPM	-0.21	-0.12	**-0.25****	-0.05	0.13	-0.22	72
		(-1.63)	(-0.82)	(-2.11)	(-0.39)	(0.99)	(-1.56)	
	FF5	-0.02	0.15	-0.13	0.08	0.23	-0.02	72
		(-0.12)	(1.00)	(-1.00)	(0.49)	(1.51)	(-0.15)	
E	CAPM	-0.21	**-0.33***	**-0.28****	**-0.22***	-0.15	0.04	72
		(-1.49)	(-1.99)	(-2.12)	(-1.69)	(-1.06)	(0.25)	
	FF5	0.03	-0.01	-0.10	0.02	0.14	**0.35***	72
		(0.23)	(-0.05)	(-0.76)	(0.15)	(0.92)	(1.77)	
S	CAPM	-0.12	**-0.27***	**-0.25***	-0.08	-0.08	**-0.25***	72
		(-0.79)	(-1.83)	(-1.99)	(-0.53)	(-0.64)	(-1.71)	
	FF5	-0.13	-0.19	-0.20	-0.10	**-0.26***	-0.20	72
		(-0.79)	(-1.19)	(-1.44)	(-0.57)	(-1.89)	(-1.18)	
G	CAPM	0.03	0.12	0.00	0.16	-0.05	-0.04	72
		(0.23)	(0.90)	(-0.03)	(1.06)	(-0.35)	(-0.25)	
	FF5	**0.26***	**0.30***	0.20	**0.33****	0.07	0.13	72
		(1.83)	(1.98)	(1.39)	(2.50)	(0.49)	(0.72)	