Sustainability ratings, equity portfolio performance, and factor models: Evidence from a multi-specification approach[a]

Working Paper – 20 March 2026

Jan Heldmann*,[b]
Department of Business Administration
University of Bayreuth
Bayreuth, Germany
Tel: +49 921 55 6258
Email: jan.heldmann@uni-bayreuth.de

Huong Dang[c]
Department of Economics and Finance
University of Canterbury
Christchurch, New Zealand
Tel: +64 3 3694059
Email: huong.dang@canterbury.ac.nz

Manuel Brinkmann[b]
Department of Business Administration
University of Bayreuth
Bayreuth, Germany
Tel: +49 921 55 6274
Email: manuel.brinkmann@uni-bayreuth.de

* Corresponding Author

Acknowledgements

[a] The authors would like to thank Morningstar Sustainalytics for helping us obtain complete rating data for S&P 500 constituents, and Peter Schmidt for advising us on the sample construction in Gibson-Brandon, Krueger, and Schmidt (2022).

[b] Jan Heldmann and Manuel Brinkmann are grateful to the Club of Finance e.V. for their generous donation to the Chair of Finance and Banking for the purchase of MSCI rating data, and Klaus Schäfer as well as the participants of 2024 International Doctoral Seminar in Banking and Finance (Liechtenstein) for their helpful comments.

[c] Huong Dang would like to thank Josh Lawson and Tong (James) Zhu for their timely research assistance.

Sustainability ratings, equity portfolio performance, and factor models: Evidence from a multi-specification approach

Abstract

Does sustainability rating information improve portfolio performance or enhance factor models? Using four ratings (ESG, Environment, Social, Governance) from five providers (Refinitiv, MSCI, RobecoSAM, Bloomberg, Sustainalytics) and a multi-specification method to create rating-sorted portfolios and factors for U.S. equities, we find little evidence of positive results from July 2003 to June 2024. Performance advantages are rare, factor model improvements are marginal, and results are highly sensitive to specification choices. Replicating three highly cited studies with our multi-specification method, we find few robust positive outcomes, particularly in the out-of-sample period.

Keywords: ESG investing; sustainability ratings; asset pricing; factor models; portfolio performance

JEL: G11, G12, M14

1 Introduction

Over the past decade, public awareness of Environmental (ENV), Social (SOC), and Governance (GOV), or combined ESG practices in financial markets has grown significantly with over $30 trillion invested globally in sustainable assets, according to the Global Sustainable Investment Alliance (2023). ESG information has been increasingly incorporated into the decision-making process of investors and asset managers. Among institutional investors, 88% report that they have increased their use of ESG information over the last year to make sustainable investment decisions (Bell and Taylor, 2024). For asset managers, over 80% state that the importance of ESG considerations has risen over the last year (Index Industry Association, 2023). Despite the increasing popularity and widespread use of ESG information in investment decisions, it remains controversial whether ESG-focused portfolios (Liang and Renneboog, 2021; Coqueret, 2022) and ESG factors (Bax et al., 2024) generate positive returns.

The latest large-scale meta-study conducted by Atz et al. (2023) analyzes over 1,100 papers and 27 meta-studies.[1] Overall, they find that ESG investments perform financially on par with conventional investments, with one third exhibiting better performance. Among 89 investor-focused studies on portfolio management strategies, ESG integration appeared in 34 studies. Atz et al. (2023) define ESG integration as a strategy which "integrates ESG analysis into fundamental research and portfolio construction [...], for example, by constructing an ESG factor." Of these 34 studies, 59% report a positive, 38% a neutral, and only 3% a negative effect.

Empirical studies on risk-adjusted returns of portfolios or factors constructed based on sustainability ratings,[2] such as Pástor et al. (2022), Díaz et al. (2021), Madhavan et al. (2021), Maiti (2021), Khan (2019), and Pollard et al. (2018), find evidence supporting a positive (adjusted) performance. On the other hand, Ciciretti et al. (2023) document negative outcomes while Hübel and Scholz (2020) and Alves et al. (2025) report largely neutral results or little evidence of systematic outperformance. Others, such as Dobrick et al. (2025), Naffa and Fain (2020, 2022), and Halbritter and Dorfleitner (2015), also observe neutral results.

Similarly, the literature is rather mixed on whether factors created based on sustainability ratings can improve the explanatory power of traditional asset pricing models.[3]

[1] Earlier large-scale meta-studies include Margolis et al. (2009), Friede et al. (2015), and Busch and Friede (2018).

[2] This study examines four distinct rating series (types): The three pillar scores (Environmental, Social, Governance) and the combined ESG rating score as reported by a rating provider. To avoid implying that our analysis focuses only on the combined measure, we use the umbrella term *sustainability ratings* to refer to any of these four rating scores. Where a cited source uses the label "ESG" we keep the original wording in the citation.

[3] For a detailed review of ESG factor attribution to portfolio returns within the Fama–French framework, see Kumar (2023).

Dobrick et al. (2025), Díaz et al. (2021), Hübel and Scholz (2020), Maiti (2021), and Jin (2018) show that adding sustainability rating factor(s) enhances the explanatory power of various Fama-French factor models. In contrast, Naffa and Fain (2022) and Xiao et al. (2013) report that including a sustainability rating factor does not significantly improve the ability of various Fama-French factor models to explain the cross-section of returns.

Two recent studies, Pedersen et al. (2021) and Pástor et al. (2021), propose theoretical frameworks that attempt to reconcile the mixed empirical findings discussed above. Pedersen et al. (2021) suggest that ESG characteristics provide insights into firm performance while also reflecting investors' preferences. They introduce the concept of the ESG-efficient frontier, which illustrates the trade-off between portfolio returns and ESG characteristics. Their framework suggests that ESG scores can lead to an increase or decrease in required returns depending on the type of investor as well as the informational value of ESG scores. Investors who prioritize ESG factors can align their portfolios with their ethical goals, however, this may come with lower returns as they do not prioritize the best possible risk-return trade-off. The second study, Pástor et al. (2021), suggests that the outperformance of assets with good ESG metrics hinges on the *taste* of investors and the need to hedge climate risks.[4] In equilibrium, stocks with better ESG metrics have a lower expected return compared to those with weaker ESG metrics because they are more in demand by ESG-motivated investors, and they additionally provide a climate risk hedge.[5] However, if a positive shock to ESG preferences or demand from investors occurs, green stocks temporarily outperform their conventional counterparts, although their expected returns decrease further.

While the above theoretical frameworks offer possible explanations for the mixed empirical results, a major source of performance variation lies in the various methods used to construct portfolios and factors with sustainability ratings. Each construction step adds potential decision choices that can materially alter the final portfolio and/or factor performance results (Walter et al., 2024; Beyer and Bauckloh, 2024; Henriquez-Salman, 2025; Cakici et al., 2025). We will outline in the discussion below how different studies take different decisions in constructing their ESG portfolios and/or factors.

First, studies differ in terms of the analysis universe including geographic coverage (region, country or index) and stock exclusions. Some studies exclude companies based on a

[4] See also Fama and French (2007) who develop a model where asset prices can be affected by disagreement and tastes for consumption-good assets.

[5] Avramov et al. (2022) extend Pástor et al.'s (2021) idea to reconcile mixed ESG portfolio return evidence by introducing ESG rating uncertainty. When ESG rating disagreement is low, brown (low rating) stocks outperform green (high rating) stocks; however, when ESG rating uncertainty rises, the performance gap shrinks or disappears.

stock price or market cap threshold, negative book value or earnings, or certain industries (Dobrick et al., 2025; Naffa and Fain, 2022; Lioui and Tarelli, 2022; Naffa and Fain, 2020). Others do not explicitly state how restricted their samples are (Pástor et al., 2022; Avramov et al., 2022; Díaz et al., 2021; Maiti, 2021).

Second, empirical studies differ in terms of the analysis periods. Díaz et al.'s (2021) study is limited to a few months while other studies cover a relatively long period, for example, five years (Naffa and Fain, 2022; Nsibande and Sebastian, 2023), eight years (Maiti, 2021; Gibson-Brandon et al., 2021), or longer than 10 years (Dobrick et al., 2025; Ciciretti et al., 2023; Hübel and Scholz, 2020; Pástor et al., 2022; Avramov et al., 2022; Pollard et al., 2018). It is worth emphasizing that not only the duration but also the economic, political, and social conditions during which a study is conducted affect the performance outcomes.

Third, the choice of the sustainability rating provider(s) matters.[6] Some studies use ratings assigned by one provider, such as MSCI (Pástor et al., 2022; Pollard et al., 2018), Refinitiv (Hübel and Scholz, 2020; Nsibande and Sebastian, 2023), Sustainalytics (Naffa and Fain, 2022; Díaz et al., 2021), or Bloomberg (Maiti, 2021), while others utilize ratings assigned by various providers (Ciciretti et al., 2023; Avramov et al., 2022; Gibson-Brandon et al., 2021; Halbritter and Dorfleitner, 2015). Furthermore, studies also differ in the ratings they analyze. Some use all four sustainability rating categories ESG, ENV, SOC, GOV (Dobrick et al., 2025; Gibson-Brandon et al., 2021; Díaz et al., 2021), others focus exclusively on the ESG rating (Ciciretti et al., 2023; Avramov et al., 2022; Naffa and Fain, 2022) or ENV rating (Pástor et al., 2022). Alternatively, Nsibande and Sebastian (2023) use Refinitiv's combined ESG score[7] while Hübel and Scholz (2020) utilize ESG, ENV, SOC separately.

Fourth, some studies adopt the Fama-French method in sorting and forming their factor portfolios[8] (Dobrick et al., 2025; Ciciretti et al., 2023; Lioui and Tarelli, 2022; Maiti, 2021), while others depart from this traditional approach (Pástor et al., 2022; Díaz et al., 2021; Gibson-Brandon et al., 2021).

Fifth, the weighting scheme used to create portfolios or factors varies across studies. Most use a value-weighting approach (Dobrick et al., 2025; Pástor et al., 2022; Avramov et al., 2022; Lioui and Tarelli, 2022; Díaz et al., 2021; Maiti, 2021; Hübel and Scholz, 2020), although

[6] Numerous studies show that sustainability rating scores vary widely across rating providers (for example, Berg et al., 2022; Gibson-Brandon et al., 2021).

[7] The combined ESG score from Refinitiv is based on the ESG score and the ESG controversy score.

[8] Fama and French (1993) form six portfolios for their factors independently based on two groups of big- and small-sized portfolios with a 50:50 split and based on another characteristic with a 30:40:30 split.

a few employ equal-weighting (Ciciretti et al., 2023; Gibson-Brandon et al., 2021; Pollard et al., 2018). There are also variations in the testing portfolios. For example, some utilize value-weighted decile portfolios (Dobrick et al., 2025), while others use equally weighted quintile testing portfolios (Maiti, 2021).

Finally, several studies additionally test whether their constructed factors could improve Fama-French factor models. However, most studies, except Hübel and Scholz (2020), do not use the comprehensive Fama-French (2018) six-factor (FF6) model as their base model. Instead, Dobrick et al. (2025) utilize the Carhart (1997) four-factor (C4) model and the Fama-French (2015) five-factor (FF5) model. Nsibande and Sebastian (2023) use the FF5 model while Maiti (2021) employs the Fama-French (1993) three-factor (FF3) model as the base model.

The methodological ambiguity discussed above creates a specification lottery where results depend as much on design decisions as on the underlying relationship between sustainability ratings and (risk-adjusted) returns. Motivated by the methodological ambiguity in the related literature, this study uses a multi-specification approach to construct sustainability rating-sorted portfolios and rating factors. Our study is most closely related to Henriquez-Salman (2025) who also quantifies methodological uncertainty in sustainability rating-related portfolio construction. However, our study markedly differs in data coverage and methodological design. Our study analyzes a broader set comprising ratings issued by the five most important rating providers (Refinitiv, MSCI, Bloomberg, RobecoSAM, Sustainalytics) over a longer period of time (2003–2024). For each rater, we utilize rating data from the first year when it was available. Our 21-year rating data of 4,481 U.S. firms accounts for evolving provider coverage with Refinitiv being the sole rating provider from 2002 to 2006, followed by MSCI in 2007, Sustainalytics in 2009, Bloomberg in 2015, and RobecoSAM in 2016. On the methodological side, we adapt methodological choices to reflect ESG data publication frequencies. Sustainability rating scores are often based on firms' annual reports and are fully updated only with a delay. For instance, ratings reflecting fiscal-year 2024 information generally become available during 2025. This timing makes annual rebalancing with appropriate lags (the Fama-French convention) more suitable than the monthly rebalancing with one-month lags used in Henriquez-Salman (2025). Our approach systematically utilizes a variety of specifications regarding filter criteria (firm size, stock price), fundamental thresholds (earnings, book value), special sector exclusions (Financials, Utilities), sorting procedure (independent, dependent), exchanges used for size breakpoints (NYSE only vs. all exchanges, namely NYSE, NASDAQ, and AMEX), and weighting schemes (value-weighted, equal-weighted). Relative to Henriquez-Salman (2025), we incorporate additional decision nodes concerning stock price and size filters, negative earnings[9]/book value, and the exchanges used to determine breakpoints. While Henriquez-Salman (2025) limits his analysis to portfolio and factor performance, we further replicate three influential ESG studies, explore portfolio return and Sharpe ratio patterns, conduct factor spanning, examine maximum ex-post Sharpe ratio, and perform Gibbons, Ross, and Shanken (GRS) tests. Using our broad data and multi-specification design, we aim to answer the following questions:

Q1: Do portfolios and factors with higher sustainability rating scores generate a higher (risk-adjusted) return than portfolios and factors with lower sustainability ratings?

Q2: Does extending the Fama-French six-factor model with a factor sorted by a sustainability rating improve its ability to explain the cross-section of returns?

To address the above questions, we first replicate the relevant analyses in three highly cited studies that examine the risk-adjusted performance of sustainability rating-related portfolios and/or factors in the U.S., which are Avramov et al. (2022), Pástor et al. (2022), and Gibson-Brandon et al. (2021). Our replications make use of the single-specification method documented in each study and our multi-specification method outlined in the subsequent section and Figure 1. We utilize the same or the most similar sample and analysis period (based on the starting time of our data) as the respective original study, and an extended (more recent) period over which the data is available to us. Our replications find that sustainability rating-related portfolio/factor performance is sensitive to specifications, factor models, and study periods. We do not obtain robust positive performance results under alternative specifications or over an extended period.

Next, we employ our multi-specification approach to create portfolios and factors sorted by either a sustainability rating type (ENV/SOC/GOV/ESG) assigned by one of the five major rating providers (Refinitiv, MSCI, RobecoSAM, Bloomberg, and Sustainalytics) or our self-created ENV/SOC/GOV/ESG rating Disagreement/Agreement.[10] For each rating provider/ rating Disagreement/Agreement, we use the rating scores in December of year $t_{(-1)}$ to construct the portfolio in June of year t_0 for the period from July of year t_0 to June of year t_1. As our rating data, updated annually, covers the period 2002-2022, our portfolio and factor analysis

[9] This sample filter has material effects on performance. Excluding versus including firms with negative earnings, while holding other decision points constant, results in large differences in median monthly returns (across specifications) of -0.16%, -0.14%, and 0.21% for Bloomberg ESG, ENV, and GOV factors, respectively.

[10] To capture the extent of consensus and divergence among rating providers, for each rating type, we compute two metrics, namely rating Agreement and rating Disagreement for each firm-year that has the rating scores from at least three providers.

period runs from July 2003 to June 2024.

We incorporate various specifications as outlined in Figure 1 and create 128 distinct specifications of one-way-sorted decile portfolios and 512 specifications of two-way-sorted 4×4 and 2×3 portfolios based on rating and size for each rating and provider. When examining one-way- and two-way-sorted portfolios, we observe no clear and consistent pattern linking higher sustainability rating scores or rating Agreement/Disagreement scores to superior financial performance. We also observe high variability across providers, as our results vary substantially depending on both the data provider and the specific rating category employed. Our rating-sorted portfolio and factor-based tests mostly show either neutral or negative performance. Likewise, most rating factors fail to attain higher maximum ex-post Sharpe ratios when combined with a purely traditional Fama-French six-factor set. Our non-positive performance results hold across alternative analysis periods, including an all-rating provider window (July 2017 to June 2024) and a crisis window that covers the COVID-19 pandemic (January 2020 to June 2024). Moreover, adding a rating factor to the Fama-French six-factor model does not substantially improve the R^2, reduce alpha levels, or lower the GRS test rejections. Our findings, which are robust to factor configurations, sustainability rating categories, and rating providers, suggest that (i) rating factors do not significantly enhance the model's ability to explain the cross-section of returns; and (ii) rating factors *may* just offer little useful information beyond Fama-French six factors.

Our study makes five additional distinct contributions. First, we systematically replicate the relevant analyses in three influential ESG studies in- and out-of-sample by using the single-specification method documented in each study and our multi-specification method. For each replication, we create 512 study-specific factor specifications which offer a much more robust and comprehensive assessment of portfolio and/or factor performance. Second, we analyze return and Sharpe ratio patterns across single-sorted and double-sorted portfolios, as well as factors. Third, we examine whether each sustainability-rating factor is spanned by the traditional Fama-French six-factor set. Fourth, we evaluate whether each rating factor improves the maximum ex-post Sharpe ratios of the extended mean-variance-efficient portfolios. Fifth, we test whether each rating factor enhances the Fama-French six-factor model's ability to explain the cross-section of returns.

Our multi-specification approach can be extended to incorporate more decision points and more choices at each step, offering practitioners a robust framework for assessing sustainability rating-based investment performance. Our study highlights the fact that methodological uncertainty can lead to flawed investment decisions and is magnified when considering various sustainability ratings (ENV/SOC/GOV/ESG) and multiple rating providers. We document large return differences across rating factor specifications. For instance, over the study period of 21 years, the ranges of returns between the worst and the best performing specifications for Refinitiv, Sustainalytics, and our self-created rating Agreement ENV factor are respectively 100.3%, 64.3%, 61.6%. We emphasize that different decisions made at one step in the portfolio construction process can cause considerable variations in factor returns. For example, switching between value- and equal-weighting when constructing a portfolio, while holding other decision points constant, leads to sizeable differences in median monthly factor returns, of -0.15% (ESG), 0.13% (SOC), and 0.26% (ENV) for the Bloomberg rating, 0.15% (ESG) and 0.17% (ENV) for the RobecoSAM rating, -0.11% (GOV) and 0.14% (SOC) for our self-constructed rating Disagreement factor. Practitioners should cautiously examine the range of observed performance outcomes beyond that of a single-specification method. Otherwise, they would risk chasing noise, misallocating capital, and failing to meet financial and/or ESG performance targets.

The remainder of our paper is structured as follows. Section 2 describes our multi-specification method to construct portfolios and factors. Section 3 summarizes the results of our replications of three influential studies. Section 4 discusses the results of our empirical analysis that examines the performance of rating-sorted portfolios and factors using our multi-specification method and ratings from five providers. Section 5 concludes with the key findings.

2. Multi-specification portfolio construction

A central problem in portfolio construction is the numerous building options which lead to “methodological uncertainty” (Walter et al., 2024). To explore whether methodological uncertainty contributes to the mixed empirical results discussed above and to assess how sensitive portfolio performance is across building options, we outline below and summarize in Figure 1 a multi-specification approach to create sustainability rating-related portfolios.

FIGURE 1 HERE

Figure 1 is a flowchart featuring the decisions we consider at each step (node) in constructing our rating-sorted portfolios.[11] The first decision (1) relates to the exclusion of companies based on their market capitalization. Small or very small companies make up a large part of the U.S. equity market and can have a major impact on portfolio returns (Landis and

[11] Our flowchart is inspired by the procedure Walter et al. (2024, Figure 3) uses to construct standard (non-ESG) portfolios.

Skouras, 2021). However, the higher returns of these small/micro caps often cannot be realized in practice due to higher transaction costs (Fama and French, 2008). The literature therefore suggests excluding a certain proportion, for example, the smallest 20% of firms (Ince and Porter, 2006), from portfolio construction. Since we do not analyze the entire market, but pre-filter by analyzing only the stocks with a sustainability rating, we modify the first decision relative to the traditional approach. Specifically, we consider two options regarding firm size: first, excluding firms with market capitalization of less than US$ 300 million;[12] second, employing no market capitalization threshold (i.e. we consider all firms with a rating in our sample).

Our second specification point (2) concerns the exclusion of firms based on stock price. Stocks with low prices can be prone to higher transaction costs. We therefore consider two options regarding stock prices: first, excluding stocks with a price of less than US$ 5; second, applying no price exclusion (Landis and Skouras, 2021). Walter et al. (2024) state that these two price thresholds ($0 and $5) are mostly used in the literature.

The four subsequent specification points relate to the exclusion of companies based on fundamentals such as negative book value (3), negative earnings (4), or being affiliated with highly regulated sectors, namely Financial (5) and Utilities (6).[13]

The next two specification points are about how long we lag the sorting variable (7) and the frequency with which we rebalance our portfolios (8). For each rating provider, rating data is practically updated annually and the rating for year t_0 is only available in year t_1.[14] Thus, we use the rating scores in December of year $t_{(-1)}$ to construct the portfolio in June of year t_0 for the period from July of year t_0 to June of year t_1 (node 7). We rebalance our rating-sorted portfolios annually considering the annual rating data (node 8). As our rating data is available from 2002 to 2022, our analysis of portfolios and factors runs from July 2003 to June 2024.

We then consider the number of quantiles to construct a portfolio. If we single-sort, we use deciles based on rating scores (10x1, nodes 10-11). If we double-sort, we either utilize four quantiles for ratings and four quantiles for size (4x4, nodes 9-10),[15] or two 50/50 size portfolios

[12] We set the minimum threshold to $300 million as the U.S. Securities and Exchange Commission (SEC) states that a microcap typically has a market capitalization of US$250–US$300 million or less. See https://www.sec.gov/about/reports-publications/investorpubsmicrocapstock.

[13] We do not employ a stock age filter. Since our sample is restricted to firms with available sustainability ratings, very young firms are unlikely to enter our initial sample, making such a filter unnecessary.

[14] The average number of months a rating change occurs ranges from 14.7 (Refinitiv) to 18.0 (RobecoSAM).

[15] While 5x5 sorted portfolios are also commonly used in similar analyses, we observe many portfolios with no stocks in our 5x5 independently double-sorted portfolio analysis (when size and rating are heavily skewed or there

and three rating portfolios formed at the usual 30 (low)/40 (medium)/30 (high) split for the ratings as in Fama and French (1993) (nodes 9-10). The former (latter) results in sixteen 4x4 (six 2x3) portfolios. Double-sorted portfolios are either independently sorted or dependently sorted (node 11).[16] For double-sorted portfolios, the breaking points for the size portfolios can be constructed by using either only stocks traded on the New York Stock Exchange or those traded on All Exchanges in the U.S. (node 12).

The final decision is whether to value-weight or equal-weight stocks in a portfolio (node 13).[17] This choice typically has a very high influence on returns, as an equally weighted portfolio weights smaller stocks higher than a value-weighted portfolio.

The above procedure results in a total of seven (nine) different specification points for each single-sorted (double-sorted) portfolio. Utilizing this framework, we have 128 specifications of decile 10x1 portfolios and 512 specifications of two-way-sorted 4x4 and 2x3 portfolios.[18]

To assess if and how specifications matter to sustainability rating-related portfolio and factor performance, we first employ the multi-specification approach discussed above to replicate three influential studies. In the subsequent section, we summarize our replication results.

3. Replication results

We replicate the relevant analyses in three highly cited studies that examine portfolio and/or factor performance linked to (i) ESG rating and ESG rating uncertainty (Avramov et al., 2022), (ii) MSCI Environment ratings (Pástor et al., 2022), and (iii) sustainability rating Disagreement (Gibson-Brandon et al., 2021).[19] Of these three studies, Avramov et al. (2022) and Gibson-Brandon et al. (2021) utilize ratings assigned by multiple providers, and the latter also employs multiple rating categories. For each study, we first utilize their single-specification method and either the same or similar datasets to create portfolios/factors that

are ties so no stock falls into quintile intersections such as big size-low rating or small size-high rating bucket). Thus, we opted for 4x4 portfolio sorting.

[16] Dependently sorted portfolios are first sorted by size, and within each size portfolio, are then sorted by rating.

[17] We also tested a capped value-weighted specification (cap of 5%) but dropped it since the results were nearly identical to those of the uncapped value-weighted version.

[18] As depicted in Figure 1, for single-sorted, there are 7 decision points resulting in 2^7=128 specifications for each portfolio. For 4x4 and 2x3 double-sorted, there are 9 decision nodes, resulting in 2^9=512 specifications for each portfolio.

[19] As of 18 October 2025, Avramov et al. (2022), Pástor et al. (2022), and Gibson-Brandon et al. (2021) have been cited 1,293, 1,159 and 984 times, respectively. Of these three, Pástor et al. (2022) is the only study that has published its code, see https://data.mendeley.com/datasets/dnskbdnmsz/1.

cover their study period or a period close to theirs (considering our data availability). Details on our replications of the original results can be found in the respective Online Appendices.[20] Next, we employ our multi-specification approach described above to construct sustainability rating-related portfolios/factors separately for (i) their study period or a period close to theirs; and (ii) an extended (more recent) period over which the data is available to us. This allows us to explore the sensitivity of rating-sorted portfolio/factor performance to alternative specification choices and an extended period.

It is important to emphasize that our replications do *not* aim at challenging the overall contributions of these studies. Our replications instead highlight how sensitive sustainability rating-related portfolio and factor performance results are to construction approaches and sample periods. By introducing a multi-specification instead of a single-specification method employed in the original studies, we are able to assess whether the reported results are robust to variations in portfolio and factor construction or whether plausible alternative specifications would have led to different conclusions regarding portfolio and factor performance. The discussion below summarizes our single- and multi-specification replicated results for each of the three widely cited studies.

3.1. "Sustainable investing with ESG rating uncertainty" by Avramov, Cheng, Lioui, and Tarelli (*Journal of Financial Economics,* 2022)

Avramov et al. (2022) form 25 value-weighted portfolios consisting of U.S. common stocks, with ESG ratings assigned between 2002 and 2018, by sorting stocks into ESG Rating-Uncertainty quintiles, and within each quintile, sorting firms into ESG-rating quintiles. They rebalance their portfolios at the end of each year. Their primary measure is PAIR, of which ESG-Uncertainty PAIR is the standard deviation of average percentile ranks across rating pairs, and ESG-Rating PAIR is the mean of their ranks.[21] They also use a robustness measure, namely ALL, which directly considers the standard deviation (ESG-Uncertainty ALL) and average percentile ranks (ESG-Rating ALL) across six examined rating providers.

In our replication, we identify the *Low Uncertainty* quintile, and within this quintile, we focus on two value-weighted portfolios, one for the quintile with the lowest and one for the

[20] For Avramov et al. (2022), see Online Appendix A. Avramov et al. (2022) – Replication Details, for Pástor et al. (2022), see Online Appendix B. Pástor et al. (2022) – Replication Details, and for Gibson-Brandon et al. (2021), see Online Appendix C. Gibson-Brandon et al. (2021) – Replication Details.

[21] Details on Avramov et al.'s (2022) sample and method can be found in Online Appendix A.1 Sample, and Online Appendix A.2 Method, respectively.

quintile with the highest rating scores.[22] We calculate the *Low ESG Rating* minus *the High ESG Rating* (*Low LMH-R*) returns. A positive *Low LMH-R* value means that when ESG ratings are credible, because rating uncertainty is low and ESG-sensitive investors accept lower expected returns on green stocks, brown (low rating) stocks earn higher returns than green (high rating) peers (Avramov et al., 2022).

Compared with the original study, our data coverage differs modestly (see Online Appendix A.1, Table A.1, Panels A-B): MSCI KLD ESG rating data is no longer available[23] while our Sustainalytics rating observations are materially greater.[24] Due to the absence of MSCI KLD data and considering our reasonable rating coverage as of 2007, our analysis starts in 2008 instead of 2003 as in Avramov et al. (2022). Our inputs, however, are qualitatively close to those of the original study. Pairwise uncertainty and correlations, as well as the quantile distributions of ESG and ESG-uncertainty (for PAIR and ALL measures), closely track the original study's values (see Online Appendix A.2., Table A.2, Panels A-C).

With our rating data from 2007, we recreate the PAIR *Low LMH-R* returns as in Avramov et al.'s (2022) Table 2 (2003-2019; Panel A: Return & Panel B: CAPM), Table 5 (2011-2019; Panel A: Return & Panel B: CAPM), Table B.4 (2003-2019; Panel A: Carhart-4 & Panel B: FF6), Table B.5 (2011-2019; Panel A: Carhart-4 & Panel B: FF6). We also reconstruct the ALL *Low LMH-R* return as in the original study's Table B.7 (2003-2019; Panel A: Return, Panel B: CAPM, Panel C: Carhart-4, & Panel D: FF6).

Besides replicating the original study using their single-specification method (value-weighted, dependently double-sorted portfolios that are rebalanced at the end of each year), we also employ our multi-specification approach using the decision points (1)-(7), (11),[25] (13) as outlined in our Figure 1. This results in 2^9=512 specifications for the *Low LMH-R* returns for each of the PAIR and ALL measures.

Figure 2 presents boxplots featuring the distributions of our multi-specification alpha estimates, our single-specification replication estimates (black dots), and Avramov et al.'s (2022) reported estimates (red dots). As in the original study, we estimate returns and alphas

[22] The *Low Uncertainty* quintile consist of the 20% stocks with the lowest rating uncertainty. Within this quintile, there are five value-weighted portfolios ranging from *Low ESG Rating* to *High ESG Rating* (Low, 2, 3, 4, High).

[23] Avramov et al. (2022) use ESG ratings from six providers. As outlined in Online Appendix A.1, we utilize ratings provided by five raters, except for MSCI KLD which was last available in 2018. KLD was one of the earliest providers of ESG data; thus, Avramov et al.'s (2022) sample starts in 2003. KLD was acquired by MSCI in 2010. They continued to publish KLD ESG data until 2018 and then discontinued it. KLD ESG rating is different from other providers' ESG ratings as it is not numerical or scaled and it tracks binary indicators of ESG strengths and concerns across companies.

[24] Our Sustainalytics data begins in 2009, whereas Avramov et al. (2022) report data starting in 2014. Replication files (data and programming code) for their paper were neither publicly available nor obtainable upon our request.

[25] For node (11), we utilize either dependently double-sorted (as in the original study) or independently double-sorted portfolios.

using four models, namely, Return (model with constant),[26] CAPM (CAPM), Carhart four-factor (C4) model, and Fama-French six-factor (FF6) model. The red dots mark the point estimates reported by Avramov et al. (2022) using their single-specification method and their samples. The black dots are our estimates using their specification and our samples. Below each box that features the distribution of our multi-specification alphas is the proportion of 5%-significant alphas for that specification set.

FIGURE 2 HERE

Across models and analysis periods, our single-specification estimates (black dots) and our multi-specification median estimates are consistently smaller than their single specification estimates, and mostly negative. Our multi-specification estimates disperse quite widely, indicating strong sensitivity of alphas to specifications, models, and analysis periods.

In Panel A (PAIR) part (i), our median estimates for the analysis period 2008–2019 are all negative (–0.21 to –0.08), while the original study's estimates for the period 2003–2019 are all positive (0.40 to 0.59). A similar discrepancy between our all-negative median estimates and their all-positive estimates can also be observed for the ALL measure in Panel B part (i).

Panel A (PAIR) part (ii) features our closest replication that covers the same sub-period as the original study, 2011–2019. Our median alpha values are mostly negative, varying between –0.27 and 0.01, whereas the original study's estimates are mostly positive, spanning between –0.08 and 0.30. Across the four models, the number of our 5%-significant alpha estimates varies between 0% and merely 6%. Since we cover the same sub-period in Panel A part (ii), deviations between our and their results can be attributed to portfolio specifications, aside from sample components.[27] For the ALL measure in Panel B part (ii), our median estimates for the period 2011–2019 are consistently smaller than their all-positive estimates for the period 2003-2019.[28]

Extending our sample period to 2008-2023 (part (iii)), our median estimates remain negative, ranging between −0.33 and −0.19 for PAIR in Panel A, and between –0.38 and –0.24 for ALL in Panel B.

Across our three analysis periods, our lower and upper values span from negative to positive values, for parts (i)-(iii) in Panel A (–0.5 to 0.51; –0.63 to 0.47; –1.12 to 0.6) and Panel

[26] *Return (model with constant)* refers to an intercept-only regression $R_t = \alpha + \varepsilon_t$, where the estimated α equals the average excess return. Testing whether $\alpha = 0$ is equivalent to a standard T-test of mean excess returns.

[27] Our sample does not include MSCI KLD ESG ratings, and we have more Sustainalytics ratings than the original study's sample. Thus, sample components (ratings assigned by various providers) may contribute to deviations between our and their estimates.

[28] Avramov et al. (2022) do not create a sub-period sample over 2011-2019 for the ALL measure. Thus, we compare our replicated result over 2011-2019 with the original study's reported result over 2003-2019.

B (–0.85 to 0.35; –0.61 to 0.52; –1.36 to 0.43), suggesting that alpha estimates are highly sensitive to portfolio specifications, factor models, and analysis periods. Plausible specifications and factor models can generate either positive alpha estimates close to the original study's results or large negative values, and this applies regardless of whether our study period is shorter than (part (i)), the same as (part (ii)) or extended from (part (iii)) the original study's. In parts (i) and (ii) of both Panel A (PAIR) and Panel B (ALL) where our analyses end in 2019, as the original study's do, the number of our 5%-significant alpha estimates for each model, except one, is less than 10%. Overall, in contrast to the positive *Low LMH-R* return and alphas reported by Avramov et al. (2022), our 512-specification replications for PAIR and ALL measures across three periods that cover various economic, political, and social conditions (2008-2019, 2011-2019, and 2008-2023) generate predominantly negative and rarely significant alphas. We note that aside from specification choices, modestly different sample components (in terms of rating providers) add to the discrepancy between our and their estimates, which suggests that performance outcomes vary across sets of rating providers.

3.2. "Dissecting green returns" by Pástor, Stambaugh, and Taylor (*Journal of Financial Economics,* 2022)

Pástor et al. (2022) construct monthly value-weighted green and brown portfolios consisting of common U.S. stocks over the period November 2012-December 2020. MSCI environmental ratings (environmental pillar score and pillar weight) are used to compute the environmental score of each firm in their sample.[29] We recreate their Figure 3 and Table 3. Detailed results of our single-specification replication can be found in Online Appendix B.3, Online Appendix Table B.1, and Online Appendix Figure B.1. We summarize below our replication results using their published code which corresponds to their single-specification method, and our multi-specification method for their exact study period and an extended period ending in June 2024.

3.2.1 Return on value-weighted green and brown portfolios (Figure 3, p. 410)

We first utilize Pástor et al.'s (2022) single-specification method and carry out the analysis corresponding to their reported Figure 3 which, for ease of convenience, is presented in our Online Appendix Figure B.1, Panel A. Our replicated cumulative returns on the value-weighted green and brown portfolios over their exact study period November 2012-December

[29] Most ratings are constant for one year or longer, even though a provider may provide monthly rating data. Pástor et al.'s (2022) MSCI environmental rating data ends in March 2020, but they conduct their analysis up to December 2020 by looking back up to 12 months when computing the environmental score of each firm. Details on their sample and method can be found in Online Appendix B.1 Sample, and Online Appendix B.2 Method, respectively.

2020 is featured in Online Appendix Figure B.1, Panel B. We obtain a cumulative return difference (CRD) of 180.1% and a monthly return spread (Green minus Brown, GMB) of 66.1 basis points (bps), which are very close to their CRD of 173.6% (171%) and GMB of 65 bps (62.1 bps) reported in their article (online[30]).

Over the extended period November 2012-June 2024, using their single-specification method, we obtain a CRD of 210% (Appendix A, Panel A).

We then employ our multi-specification approach to construct green and brown portfolios. We use the decision points (1)-(6), modified (7), modified (10), and (13) as in our Figure 1. It is worth emphasizing that the authors use the most recent environmental scores instead of adopting the Fama-French (FF) lagged scores (node 7, our Figure 1) and they do not use the FF common split of 30/40/30 (node 10, our Figure 1) but a three times 1/3 break instead. Thus, we have two "modified" decision points (either the most recent scores or FF lagged scores, and either a three times 1/3 break or FF 30/40/30 split) totaling nine, resulting in 2^9 =512 specifications.

Appendix A, Panel B depicts the ranges of cumulative returns for our 512 specifications of green and 512 specifications of brown portfolios over the period November 2012-June 2024. It is noticeable that the 5% to 95% (20% to 80%) cumulative return ranges of both portfolios overlap during 2016-late 2018 and from 2021 (early 2022) onwards. If we end our analysis in December 2020 (as in the original study), our CRDs were all positive, ranging between 61.5% and 320.8%, and the median (mean) CRD was 163.6% (163%), which is close to their single-specification CRD of 173%. However, if we extend the study period to June 2024, the result is less in favor of the green portfolio as CRDs range widely between -78.4% and 268% and about 11% of specifications have either negligible positive or negative CRDs. The median (mean) CRD was more modest, at 136% (109.9%), which is much lower than 210% (using the authors' method and the extended study period, see Appendix A, Panel A). That is, their single-specification would-be-CRD of 210% for the extended period is close to our 85% quantile with a CRD of 211.5%.

3.2.2. Green minus Brown (GMB) spread performance (Table 3, p. 411)

We replicate Pástor et al.'s (2022) analysis that examines whether the GMB spreads can be well explained by factors in traditional asset pricing models (Table 3 in their article). Specifically, we repeat the analysis corresponding to models in columns 1 (Return), 2 (CAPM),

[30] Updated data can be found on Pástor's webpage: https://faculty.chicagobooth.edu/lubos-pastor/data

3 (Fama-French three-factor (FF3) model), 4 (Carhart four-factor (C4) model), and 6 (Fama-French five-factor (FF5) model).

Online Appendix Table B.1 reports the coefficient estimates and T-values (in bold if significant) of the returns (alphas) from their analysis, our single-specification replications ending in December 2020 and in June 2024. The alphas are no longer statistically significant at the 5% level, except for FF5, when the extended period November 2012-June 2024 is analyzed.

We then employ our multi-specification approach and estimate 512 specifications of each model for the period ending in December 2020. We show the range of alphas and report the statistics of alphas in Figure 3, Panel A. In each of the four estimated models, not 100% of specifications generate statistically significant alpha. About 1%, 17%, 20%, and 15% of alphas derived from CAPM, FF3, C4, and FF5 models are not significant at the 5% level.

FIGURE 3 HERE

Next, we estimate 512 specifications of each model for the extended period ending in June 2024. The range of alphas and the statistics of alphas are depicted in Figure 3, Panel B. It is noticeable that the median and mean alphas derived from all models are much smaller in Panel B (extended period) than in Panel A (original study's period), and most alphas are no longer statistically significant at the 5% level. Only 0%, 14%, 12% and 27% of alphas obtained from the CAPM, FF3, C4 and FF5 models are significant. The result of our replication using Pástor et al.'s (2022) single-specification method over the extended period and the result of our multi-specification analysis emphasize that the study period and the specifications used to construct portfolios matter.

Overall, compared with Pástor et al. (2022), our single-specification replication confirms their green outperformance in the original time window. However, our multi-specification analysis shows that the result is sensitive to specifications (and factor models), and when extended to 2024, the positive performance largely fades with most alphas being no longer significant. We therefore cautiously conclude that their findings are not robust out of sample and may not hold under reasonable alternative specifications.

3.3. "ESG Rating Disagreement and Stock Returns" by Gibson-Brandon, Krueger, and Schmidt (*Financial Analysts Journal*, 2021)

Gibson-Brandon et al. (2021) create quintile industry-adjusted rating disagreement-sorted portfolios for S&P 500 stocks between 2010 and 2017. They adjust for the respective firm's industry each month by subtracting the Fama–French 12-industry mean. Firm-level rating disagreement is measured as the standard deviation of percentile-ranked ESG/ENV/SOC/GOV scores across seven examined providers. The single-sorted quintile portfolios are rebalanced in January each year based on December's rating disagreement values. The authors focus on the equally weighted Low (Q1), High (Q5), and High–Low (H–L) portfolios. Portfolio and factor performance is evaluated using five models: Return, CAPM, FF3, Carhart, and FF5.

We recreate Gibson-Brandon et al.'s (2021) sample of S&P 500 constituents[31] from 2010 and utilize available sustainability ratings issued by four major providers (Refinitiv, Bloomberg, Sustainalytics, MSCI). According to the authors, these are the most important rating providers (p. 107). Our rating data coverage and cross-provider correlations closely track the original study (Online Appendix C, Tables C.1-C.3).[32] Our industry-adjusted rating disagreement distributions also align closely with the corrected series supplied by the authors (Online Appendix C, Table C.4).

3.3.1. Single-specification replication: Portfolio Sorts on Industry-Adjusted Rating Disagreement (Table 5, p. 116)

Using the authors' single-specification method (percentile ranks, monthly industry adjustment, yearly rebalancing, quintile single-sorted, equal weights) we replicate portfolio returns and factor alphas and report the results in Online Appendix Table C.5. We focus on the High minus Low (H–L) portfolios in our brief discussion below. For the period 2010–2017 (see "Replication" row), our ESG returns and alphas (Panel A) are positive and significant, which closely matches with the original study's. ENV (Panel B) estimates are positive, however, most alphas are insignificant, which is inconsistent with the original study's results. SOC (Panel C) and GOV (Panel D) estimates are insignificant. The inconsistences (for ENV and SOC estimates) may be due to a smaller rating provider set that we employ compared to the original study.[33] For the extended period 2010–2023 (see "Replication+" row, Panels A-D), we observe insignificant estimates across models and rating dimensions. For ESG and ENV, our H-L portfolio returns and risk-adjusted alphas are smaller than the original study's.

[31] We obtain monthly S&P 500 constituents from Refinitiv's Datastream.

[32] Details on the Gibson-Brandon et al. (2021) sample and method can be found in Online Appendix C.1 Sample, and Online Appendix C.2 Method, respectively.

[33] As indicated in Online Appendix C.1., restricted by data availability, our sample does not include ratings from MSCI KLD, which were last available in 2018, and the two less popular raters, namely, FTSE and Inrate.

3.3.2. Multi-specification replication: Portfolio Sorts on Industry-Adjusted Rating Disagreement (Table 5, p. 116)

We construct rating disagreement portfolios using our multi-specification framework, specifically decision points (1)-(6), modified (7), modified (10), and (13), as depicted in Figure 1. It is worth emphasizing that the authors utilize quintile sorts instead of FF 30/40/30 breaks (node 10, our Figure 1) and form portfolios each January based on December's rating disagreement values instead of using an FF lag (node 7, our Figure 1). Thus, we have two "modified" decision points (either December's scores or FF lagged scores, and either quintile single-sorted or FF 30/40/30 split single-sorted) totaling nine, resulting in 2^9 =512 specifications.

Figure 4 presents the distribution of 5%-significant alpha estimates for the original study period 2010-2017 (Panel A) and the extended period 2010-2023 (Panel B). During the original period 2010-2017, a modest number of alphas for ENV (Panel A.2, 13.3%-25.4%) and SOC (Panel A.3, 2%-16.8%), and a small number for ESG (Panel A.1, 4.3%-9.4%) and GOV (Panel A.4, 1.2%-9.8%) are statistically significant at the 5% level. Untabulated analysis reveals that the median (mean) alphas across all rating categories and factor models are positive, ranging between 0.7% and 15.5% (0.4% and 14.7%). Over the extended period 2010-2023, 6.3%-25% of alphas for SOC (Panel B.2), 3.1%-22.3% of alphas for GOV (Panel B.4), and merely less than 2.5% of alphas for ESG (Panel B.1) and ENV (Panel B.2) are statistically significant. The H-L portfolio performance outcomes vary across rating dimensions, with the mean and median alphas being positive for SOC and GOV but negative for ESG and ENV (untabulated).

FIGURE 4 HERE

Overall, compared with Gibson-Brandon et al. (2021), our single-specification replication reproduces their H-L positive premiums for ESG (significant) and GOV (insignificant) relatively well in the original study period. However, our multi-specification replications over the original and extended windows 2010-2017 and 2010-2023 show that the majority of alphas are statistically insignificant. Over the extended period ending in 2023, the positive mean/median premiums largely fade and become negative for ESG and ENV. We should note that our sample includes ratings provided by the four most important raters while the original study utilizes ratings assigned by seven providers. Thus, sample differences and portfolio specifications *may* jointly explain discrepancies between our multi-specification replications and the authors' single-specification results in the original window.

In short, our replications of the three influential studies reveal that sustainability rating-sorted portfolio/factor performance is sensitive to portfolio specifications and study periods, and varies across rating provider sets. Of the three highly cited studies we replicated, Gibson-Brandon et al. (2021) is the only one that utilizes multiple rating categories and multiple rating providers; however, it uses only one measure (rating disagreement) to form portfolios. Avramov et al. (2022), while employing multiple rating providers, use only one rating type (ESG) to create two measures (ESG rating and ESG rating uncertainty PAIR/ALL). Pástor et al. (2022) only make use of one rating pillar (ENV) and one rating provider (MSCI). In terms of study periods, except for Avramov et al. (2022) which ended in December 2020, the other two studies ended a few years before the global outbreak of the COVID-19 pandemic. The empirical setups of these three influential studies and our replicated results motivate us to conduct a more comprehensive study that covers a long period of time including the COVID-19 pandemic and utilizes a multi-specification framework to construct portfolios/factors based on various rating categories provided by the most important raters. In the subsequent section, we will summarize the results of our self-created sustainability rating-based portfolios and factor performance analyses.

4. ESG portfolio and factor performance using the multi-specification approach

In this section, we will discuss the results of our empirical analysis that examines the performance of rating-sorted portfolios and sorted factors utilizing the multi-specification approach depicted in Figure 1 and sustainability ratings assigned by five popular providers.

4.1. Sustainability rating data

For our empirical analysis, we collect monthly stock returns and market equity from CRSP for all common U.S. stocks (with a share code of 10 or 11) listed on NASDAQ, NYSE or AMEX from January 2002 to June 2024. Thereafter, we select firms with annual sustainability ratings, specifically, Environmental (ENV), Social (SOC), Governance (GOV), and ESG rating scores from at least one of the five popular providers: Refinitiv (REF), MSCI, Sustainalytics (SUS), RobecoSAM (ROB), and Bloomberg (BB).[34] Our sustainability rating data start with Refinitiv in 2002, followed by MSCI in 2007, Sustainalytics in 2009, Bloomberg in 2015, and RobecoSAM in 2016. Considering rating data availability, our main rating sample covers the period January 2002-December 2022 and consists of 4,481 unique firms, each of which had at least one sustainability rating issued by one of the five providers.

[34] We use Bloomberg sustainability ratings instead of Bloomberg disclosure scores as the latter is solely based on a firm’s disclosure level.

Except for Sustainalytics, all rating providers maintained a consistent method during our study period. For Sustainalytics, prior to 2019, a high ENV/SOC/GOV/ESG rating score indicated an asset with "good" values. From 2019 onwards, a high Sustainalytics score indicates an asset with high risk. Thus, we invert Sustainalytics rating scores assigned from 2019 to 2022 to make them consistent with Sustainalytics ratings issued during the preceding period[35] and ratings issued by other providers. In our analysis, a portfolio formed with a low Sustainalytics score, for example, a low ESG rating, can be interpreted as including ESG-risky assets, whereas a portfolio constructed with a high Sustainalytics ESG score consists of ESG-valued assets.

For each rating score *s* (ENV/SOC/GOV/ESG), we create two measures, namely, Agreement (AGR) and Disagreement (DIS). We follow the literature in calculating the average and standard deviation of our sustainability ratings (Avramov et al., 2022; Alves et al., 2025).[36] First, we normalize the rating score for firm f issued by provider p at time *t*, $s_{f,p,t}$, as follows:

$$ns_{f,p,t} = \frac{(s_{f,p,t} - s_{min,p,t})}{(s_{max,p,t} - s_{min,p,t})} * 100 \tag{1}$$

Where:

$ns_{f,p,t}$ is the normalised rating score assigned for firm *f* at time *t* by rating provider p.

$s_{min,p,t}$ ($s_{max,p,t}$) is the lowest (highest) rating score assigned by provider p at time *t*.

The normalized score $ns_{f,p,t}$ ranges between 0 – 100. The Agreement score $ns_{agree,f,t}$ for firm *f* at time *t* is the mean of the normalized scores $ns_{f,p,t}$ provided by at least three of the five providers *p*.[37]

$$ns_{agree,f,t} = \begin{cases} \frac{1}{N} \sum_{p=1}^{N} ns_{f,p,t} & if\ N \geq 3 \\ NA & if\ N < 3 \end{cases} \tag{2}$$

The Disagreement score $ns_{disagree,f,t}$ for firm *f* at time *t* is computed as the standard deviation of the normalized scores $ns_{f,p,t}$:[38]

[35] We invert by calculating (100 minus the Sustainalytics rating score). Thus, a score of 10 indicating low risk in the new system now would get an inverted high score of 90, which represents a good ENV/SOC/GOV/ESG value.

[36] Avramov et al. (2022) and Alves et al. (2025) measure agreement and disagreement (uncertainty) based on monthly percentile rankings relative to other stocks. We differ in that we calculate rating agreement and disagreement scores from annually normalized provider scores. Beyond the simple average, Berg et al. (2024) compare several aggregation methods and caution that no single aggregator is probably superior. Therefore, we use the simple average as it is transparent and easily implemented.

[37] Avramov et al. (2022) require ratings from at least two agencies to compute the rating average and uncertainty while Gibson-Brandon et al. (2021) calculate rating disagreement using ratings assigned by at least three providers.

[38] We correct the denominator to *(N−1)* to obtain an unbiased estimator of the standard deviation of ratings. This adjustment does not materially affect portfolio construction or ranking outcomes.

$$ns_{disagree,f,t} = \begin{cases} \sqrt{\frac{1}{N-1}\sum_{p=1}^{N}\left(ns_{f,p,t} - ns_{agree,f,t}\right)^2} & if\ N \geq 3 \\ NA & if\ N < 3 \end{cases} \tag{3}$$

Online Appendix Table D.1 includes statistics of normalized rating scores in our universe of rated firms.[39] The average and median ratings vary across rating categories and rating providers. For each rating category (ESG, ENV, SOC, GOV) assigned by a rating provider or ESG/ENV/SOC/GOV Agreement/Disagreement, we use the procedure outlined in Figure 1 to construct portfolios. Online Appendix Table D.2 shows the percentage of firms being excluded at each decision point individually in Figure 1.

4.2 Rating-sorted portfolio performance

To examine whether portfolios with higher sustainability rating scores generate higher risk-adjusted returns than those with lower corresponding scores, we first create five one-way-sorted rating portfolios and compare their excess returns and Sharpe ratios using boxplots. If there is a positive relationship between sustainability rating scores and financial performance, we should observe an increasing median excess return and Sharpe ratio value from portfolio 1 with the lowest-rated companies to portfolio 5 with the highest-rated firms.

Appendix B Panel A shows the excess return boxplots of rating-sorted quintile portfolios for each rating type (ENV/SOC/GOV/ESG) and each rating provider/our self-constructed rating Agreement/Disagreement. Across all providers and rating types, the relationship between rating scores and excess returns is inconsistent. Median excess returns fluctuate widely across portfolios, and some ratings exhibit a downward-sloping trend, for example, AGR's ESG and SUS's GOV. Furthermore, the variation in excess returns, indicated by the box containing 50% of all values, differs significantly across portfolios. High differences in interquartile range values are also present across rating types and providers.

A potential reason for the inconsistent relationship between rating scores and excess returns could be different volatility levels for low-rated and high-rated portfolios. Thus, we investigate the Sharpe ratios of rating-sorted portfolios for each provider and each rating type, and report the results in Appendix B Panel B. We generally observe more favorable Sharpe ratios for higher-rated portfolios. For example, while AGR's ESG is downward-sloping with median excess returns declining from 1.29 (portfolio 1) to 1.14 (portfolio 5) (Panel A), its median Sharpe ratios slightly increase from 0.82 (portfolio 1) to 0.92 (portfolio 5) (Panel B).

[39] Agreement and Disagreement capture the market-wide perception about a firm, thus, we compute them based on our universe of U.S. rated firms. Computing them after filtering as outlined in Figure 1 would make the measures sample-dependent (up to 2^6=64 variants), which conflicts with investors' market-wide information set.

However, across rating categories and providers, we are still not able to depict recognizable patterns along portfolios' rating ranks.

As larger companies tend to have higher rating scores[40] and smaller companies empirically tend to have higher returns, we repeat this analysis for 4x4 double-sorted portfolios based on market capitalization (size) and a rating type. We plot median excess returns (Appendix C Panel A) and Sharpe ratios (Appendix C Panel B) of our 512 specifications for each portfolio in a heat plot. Assuming a positive (negative) relationship between rating (size) and financial performance, we should observe the median excess return to increase as we move from the portfolio with the lowest scores (rating portfolio 1) to that with the highest scores (rating portfolio 4), and from the portfolio of larger companies (size portfolio 4) to the portfolio of smaller ones (size portfolio 1).

A darker shade in the heat plot indicates a higher median value and better financial performance. For REF (the earliest rating provider with the largest number of rating observations), across rating types (ENV/SOC/GOV/ESG), big size portfolios with lower rating scores perform the worst in terms of excess return, and except ENV, risk-adjusted return. The excess returns, however, differ across rating providers and rating categories. The pattern shifts a bit when we examine risk-adjusted returns, with most big size portfolios enjoying greater Sharpe ratios. However, this does *not* necessarily lead to portfolios with higher rating scores consistently delivering better risk-adjusted performance. As with the one-way-sorted quintile portfolios, it is not possible to establish a clear and consistent relationship between ratings and returns/risk-adjusted returns.

Our additional analysis of one-way-sorted quantile portfolios and 4x4 double-sorted portfolios based on size and rating over two sub-periods, an all-rating provider window, July 2017 to June 2024, and a crisis window, January 2020 to June 2024, produces qualitatively similar results with no systematic monotonicity (untabulated).

4.3. Rating factor performance

4.3.1 Sustainability rating factors

For each rating type (ENV/SOC/GOV/ESG) issued by each of the five rating providers or our self-computed rating Agreement/Disagreement, there are 512 (2^9=512) 2x3 portfolio

[40] Comparing companies in the top and bottom market-cap quartiles, we find that median normalized ESG scores are higher for large firms across all providers: +26 points for Refinitiv, +10 for MSCI, +28 for Bloomberg, +21 for Sustainalytics, and +31 for RobecoSAM.

specifications (as outlined in Figure 1) and 512 corresponding factor specifications constructed as described below.

For each specification, for example, using the MSCI ESG rating, we create the corresponding MSCI ESG factor by either dependently sorting or independently sorting stocks in the portfolios. If we dependently sort, we first sort stocks by Size into a Small and a Big portfolio based on the median market capitalization. Next, within each Size portfolio, we further group stocks into three sustainability rating portfolios based on their respective normalized rating: Low Rating (0–30%), Neutral Rating (30–70%), and High Rating (70–100%). Alternatively, we conduct two independent sorts. First, we sort stocks by Rating (and get Low/Neutral/High rating portfolios). Next, we separately sort all stocks by Size into a Small and a Big portfolio. The final portfolios we use to construct the corresponding MSCI ESG factor specification are the intersection of the Rating and Size portfolios (Small-High, Small-Low, Big-High, Big-Low). Regardless of the sorting procedure, the MSCI ESG factor is computed as the average return difference between the two High-Rating portfolios and the two Low-Rating portfolios:

$$MSCI_{ESG}\ Factor = \frac{1}{2} * (Small\text{-}High\ + Big\text{-}High\,) - \frac{1}{2} * (Small\text{-}Low\ + Big\text{-}Low\,) \quad (4)$$

Appendix D reports the differences in median monthly returns (across factor specifications) due to alternative decisions made at one point (node) at a time when constructing a portfolio, holding other specification points in Figure 1 constant. The largest shifts arise from the weighting decision and whether to include firms with negative earnings. Choosing value-weighting instead of equal-weighting (column 2) results in sizable differences in median monthly factor returns, for example, for Bloomberg rating, Panel C (+0.26% for ENV, +0.13% for SOC; -0.15% for ESG), RobecoSAM rating, Panel D (+0.17% for ENV, +0.15% for ESG), Rating Disagreement, Panel G (+0.14% for SOC, -0.11% for GOV). Similarly, excluding versus including firms with negative earnings (column 6) leads to noticeable variations in median monthly factor returns, for example, Bloomberg rating, Panel C (-0.16% for ESG, -0.14% for ENV, +0.21% for GOV). Overall, different decisions made at one step in the portfolio construction process can cause considerable variations in median factor returns.

4.3.2 Rating factor returns and risk-adjusted performance

Appendix E shows the median and range of monthly returns across specifications of factors categorized by rating providers and rating types, over time. We do not observe any clear upward trend. Instead, median factor returns either fluctuate or exhibit a general downward trend as time passes, except for REF’s SOC. Notably, some factors for SUS, ROB, BB, AGR, and DIS generated large median negative returns either during most of or throughout the study period.

Appendix F reports the statistics of cumulative returns across factor specifications, as of June 2024, categorized by rating types and rating providers. Over our study period of 21 years, for most providers, ENV factor returns vary widely across specifications. For example, the worst- and best-performing specifications of the ENV factor delivered returns between 1.7% and 102%, -58% and 6.3%, -35.2% and 26.4% for REF, SUS, and AGR, respectively. The SOC factor also experienced large variations across specifications. Its worst and best specifications generated returns of between 3.2% and 72.9%, -13.5% and 52.8% for REF and MSCI respectively. These examples emphasize that investment performance outcomes shift considerably across rating factor specifications for the same rating provider and rating dimension.

To examine whether rating factors generate positive (risk-adjusted) returns, we first examine their returns and Sharpe ratios using box plots. If companies with high rating scores perform better financially than companies with lower scores, we should expect a box in the boxplot to be in the positive range, which would mean that 75% of the specifications have positive returns/Sharpe ratios. If the box is on the zero line, we assume an overall neutral or an ambiguous effect of ratings on returns/Sharpe ratios. If the box is on the negative side, companies with lower rating scores financially outperform those with higher scores.

Figure 5, Panel A, shows the boxplots of factor returns for each provider and each rating type. Across all rating categories (ENV/SOC/GOV/ESG), REF boxes are the only ones in the positive-return region (with positive median returns). Behind REF, MSCI’s three factors (SOC/GOV/ESG) exhibit quite similar positive performance. In contrast, ROB and SUS boxes are either entirely or mostly in the negative-return territory (with negative median returns).

FIGURE 5 HERE

Panel B features the boxplots of rating factors’ risk-adjusted returns for each provider and each rating type. Similar to Panel A, REF’s factors enjoy positive Sharpe ratios across rating types. For MSCI’s (DIS’s) factors, we observe positive Sharpe ratios for three (two) rating categories. For the remaining providers (BB, ROB, SUS, AGR), Sharpe ratios are either neutral or negative and generally centered at or below zero with limited positive tails.

Untabulated analysis over the two sub-periods, July 2017–June 2024 (all-rating provider window) and January 2020–June 2024 (crisis window), reveals that sustainability rating-sorted factor returns and Sharpe ratios are similar to those of the baseline analysis (mostly neutral or negative).

4.3.3 Maximum ex-post Sharpe Ratio

The previous analysis suggests that sorted portfolios with high rating scores do not necessarily generate better excess returns or risk-adjusted returns, and most rating factors deliver negative risk-adjusted returns. To thoroughly explore the economic significance of our rating factors, we compute the maximum ex-post Sharpe ratios of the mean-variance-efficient portfolios extended by a rating factor. The base portfolio includes the six factors from Fama-French (2018), which are market excess return (R_m-R_f), Small Minus Big (*SMB*), High Minus Low (*HML*), Robust Minus Weak (*RMW*), Conservative Minus Aggressive (*CMA*), and Momentum (*MOM*), which we obtain from French's data library.[41]

Aside from the maximum ex-post Sharpe ratios, we also explore whether the bias-adjusted squared Sharpe ratio for an extended portfolio with an added rating factor is higher than that of the base portfolio. As Hanauer (2020) suggests, this is equivalent to testing whether the rating factor in the extended portfolio that is not in the base portfolio has a significant alpha when regressed on the base portfolio.[42] Since we construct 512 factors, we have, accordingly, 512 extended portfolios with an additional factor for each sustainability rating and rating provider. Figure 6 depicts the distribution of the Sharpe Ratio difference (Panel A, ΔSR) and adjusted squared Sharpe Ratio difference (Panel B, ΔaSR^2) between the extended portfolio and the base (FF6) portfolio.

FIGURE 6 HERE

Adding a rating factor leads to, at best, very modest improvements in the ex-post Sharpe ratio (Panel A), mainly for REF, MSCI, and DIS. Across all providers, median ΔSR values remain below 0.05. For BB, ROB, SUS, and AGR, there is virtually no change, while the most noticeable gain appears for DIS's ENV factor, where the upper tail of the box distribution reaches around 0.08. In general, the improvements are economically small.

[41] See https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

[42] Using the squared Sharpe ratio allows for a more straightforward and statistically rigorous comparison of models because its sampling properties are better behaved and more directly linked to mean-variance efficiency. Following Hanauer (2020), we correct each model's squared Sharpe ratio for small-sample bias under joint normality. For this purpose, we first multiply our squared Sharpe ratio by $(T - K - 2)/T$ and subtract K/T afterwards. Here, K is the number of factors while T is the number of return observations.

For the bias-adjusted squared Sharpe ratio difference (ΔaSR^2, Panel B), the picture is similar. We observe positive median ΔaSR^2 for only three cases: SOC for REF, ENV for MSCI, and ENV for DIS; however, the gains are statistically insignificant.

Over the two sub-periods, July 2017–June 2024 (all-rating provider window) and January 2020–June 2024 (crisis window), we observe qualitatively similar results (untabulated), despite possible short-sample period noises. Overall, adding a sustainability rating factor to the base portfolio including FF six factors does not materially improve the maximum ex-post Sharpe ratios.

4.3.4. Rating factor spanning

We further investigate factor spanning, which tests whether each of our rating factors is either fully or partially explained (spanned) by traditional factors used in asset pricing, specifically the factors of the FF6 model. To do so, we estimate the following time-series regression:

$$R_{t,s,p,i}^{RatingFactor} = \alpha + \beta_1\left(R_m - R_f\right)_t + \beta_2 SMB_t + \beta_3 HML_t + \beta_4 RMW_t \qquad (5)$$

$$+\beta_5 CMA_t + \beta_6 MOM_t + \epsilon_t$$

Where:

$R_{t,s,p,i}^{RatingFactor}$ is the return of the rating factor of rating score *s* (ESG, ENV, SOC, GOV) issued by a "provider" *p* (Refinitiv, MSCI, Bloomberg, RobecoSAM, Sustainalytics, or self-calculated rating Agreement/Disagreement) for each of our 512 factor specifications *i* in month *t*.

R_m- R_f, *SMB*, *HML*, *RMW*, *CMA* and *MOM* are the six factors from the FF6 model.

This test allows us to determine if a rating factor offers any independent useful information or if it is redundant when combined with the factors from the FF6 model. Additionally, we assess whether rating scores assigned by different providers yield consistent results or vary in their overlap with the FF6 factors. As we construct 512 specifications for each factor, we estimate 512 specifications for each model. Table 1 shows the median coefficient estimates of FF6 factors (columns 3-8), the percentage of specifications in which a coefficient estimate is significant at the 5% level (in parentheses below the coefficient estimate, columns 3-8), and the median R^2 (column 9).

TABLE 1 HERE

The explanatory power (median R^2) of FF6 factors (column 9) is significantly higher for some providers than for others, for example, BB (19.1% to 58.6%), Agreement (25.2% to 37.3%), ROB (25.8% to 30.9%). The median R^2 is particularly low for Disagreement (3.4% to 18.5%). Across rating providers, the GOV and ENV factors are better explained, with the respective median R^2 average values of 28.4% and 29.2%, while for the other two factors, median R^2 averages are 19.4% (ESG) and 21.5% (SOC).

The number of specifications for which an FF6 factor is significant in explaining a rating factor differs greatly across providers and rating types. For example, considering the CMA factor (column 8), for BB's GOV, 100% of its specifications are significant, whereas for SUS's ESG, none is significant.

Overall, consistent with earlier findings, we observe substantial heterogeneity across providers and rating types, and large variations in the number of significant FF6 factor loadings across specifications. Most importantly, rating factors are not uniformly or strongly spanned by FF6 factors, suggesting that rating factors may carry incremental information. This raises the question: Does adding a rating factor to the FF six-factor model improve its ability to explain the cross-section of returns?

4.3.5. Gibbons, Ross and Shanken (GRS) Test

We further employ the Gibbons, Ross, and Shanken (GRS) test, which assesses the joint hypothesis that all portfolio alphas equal zero, to investigate if a portfolio set as a whole can be better explained by the addition of a rating factor.[43] This approach is similar to Hou *et al.* (2021) testing their q-factor model against other common factor models with a large set of testing portfolios. To carry out the GRS-test, we first run the following time-series regression for each testing portfolio:

$$R_t^{TP} = \alpha + \beta_1 \left(R_m - R_f\right)_t + \beta_2 SMB_t + \beta_3 HML_t + \beta_4 RMW_t + \beta_5 CMA_t + \beta_6 MOM_t + \beta_7 R_{t,s,p,i}^{RatingFactor} + \epsilon_t \qquad (6)$$

Where:

[43] We add one rating factor at a time for three reasons. First, ESG is an aggregate of ENV, SOC, and GOV, and the four sustainability ratings are highly correlated. In our sample, the average within-provider correlations (pairwise among ESG, ENV, SOC, GOV) are quite high, for example, the mean correlation is 0.88 for ROB and 0.75 for SUS. Including all four simultaneously introduces multicollinearity that inflates standard errors and yields unstable betas, weakening inference. Second, in a GRS setting, adding correlated factors reduces degrees of freedom without matching improvements in fit, lowering power for alpha/GRS tests. Third, our test-asset sets are matched to the factor under test (e.g., ENV factor evaluated on ENV-sorted portfolios). Mixing all four rating factors would require a mixed test-asset set (ENV/SOC/GOV/ESG portfolios together), complicating interpretation of the joint alpha test. For these reasons, testing incremental models with one added factor at a time is the most straightforward and effective approach.

$R_{t,s,p,i}^{RatingFactor}$ is the return of the rating factor of rating score *s* (ESG, ENV, SOC, GOV) issued by a "provider" p (Refinitiv, MSCI, Bloomberg, RobecoSAM, Sustainalytics, or self-calculated rating Agreement/Disagreement) for each of our 512 factor specifications i in month *t*.

R_t^{TP} is the return of the testing portfolio *TP* in month *t*.

R_m-R_f, *SMB*, *HML*, *RMW*, *CMA* and *MOM* are the six factors from the FF6 model.

Our testing portfolios comprise two blocks: a sustainability set and a standard set. The sustainability set mirrors the rating dimensions we test (ESG, ENV, SOC, GOV) across all seven "providers". For each rating score we form 14 sustainability portfolio sets, seven decile (10×1) sorts on the score (one per provider) and seven Size×Score 4×4 double-sorts. For each factor, we evaluate the same specification against its matched portfolio specification.[44] The standard set has two sub-sets: (i) 193 Global-Q sets which are primarily one-way 10×1 sorts (Hou et al., 2020);[45] and (ii) seven two-way 5×5 sorts from French's data library.[46]

Table 2 features the results of the analysis that evaluates whether adding a single sustainability rating factor to the FF6 baseline model improves its ability to explain the cross-section of returns. We report differences (Δ) or percentage point differences (%Δ) relative to the FF6 baseline for matched Sustainability Sets (left block) and broad Standard Sets (right block). A negative (positive) change in the number of GRS-test rejections at the 5% level (p(GRS-T)<5%, columns 2 and 9) means a better (worse) fit. A positive (negative) change in average adjusted R^2 ($\overline{R^2}$, columns 3 and 10) implies a relative higher (lower) explanatory power, while a positive (negative) change in $\overline{R_H^2 - R_L^2}$ (columns 4 and 11) indicate more (less) dispersion across portfolios. A more negative (positive) change in number of alphas significant at the 1% (column 5 and 12) and 5% (columns 6 and 12) also indicate a better (worse) model fit. A relative negative (positive) average absolute alpha $\overline{|\alpha|}$ (columns 7 and 14) means smaller (larger) mispricing magnitudes. Finally, a positive (negative) change in $\overline{\alpha_H - \alpha_L}$ (columns 8 and 15) means that the high-minus-low alpha gap widens (shrinks), indicating a worse (better) fit.

TABLE 2 HERE

[44] This means that if a factor is created with the criteria including the minimum capitalization of US$300 million, a minimum share price of US$5, no exclusion of negative book value or negative earnings, and inclusion of financial and utility companies, the portfolio set (against which a factor is tested) is created with the same criteria.

[45] These include 42 momentum, 32 value-versus-growth, 32 investment, 46 profitability, 31 intangibles, and 10 friction sets. Some anomalies have only five or nine portfolios instead of ten. See https://global-q.org/testingportfolios.html.

[46] These include Size × (Book-to-Market, Operating Profitability, Investment, Prior 12–2 Returns); Book-to-Market × (Investment, Operating Profitability); and Operating Profitability × Investment. See https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

When we evaluate against the sustainability portfolio sets (left block), improvements are provider- and rating-specific, and generally modest. The model extended by REF ENV (Panel A) improves relative to the base model the most with a reduction in GRS test rejections of 31% (column 2), an increase of +0.03 in average explanatory power (column 3), and noticeable reductions in significant portfolio alphas at the 1% and 5% level (–2.1% and –6%, columns 5 and 6). We also observe smaller improvements in GRS test rejections for REF (Panel A) ESG (-4.5%), and GOV (-8.8%), and Sustainalytics (Panel E) ESG (-3.4%), ENV (–7.4%), SOC (–8.4%), and GOV (–11.9%). Most other cases show little improvement or even deterioration.

When we evaluate against the broad standard test assets (right block), the incremental value is negligible or negative. Changes in ΔR^2 (column 10) are essentially zero, and alpha-based metrics (columns 12, 13, and 14) alter marginally. In few cases, we observe small improvements (−0.5% to −1.5%) in the number of GRS-test rejections (column 9), but these are not accompanied by meaningful gains in R^2 (column 10) or reductions in alpha magnitudes (column 14). The metrics are even worse for several providers, for example, Bloomberg (Panel C) shows an increase in GRS test rejections of between +2% and +5%, and Sustainalytics (Panel E), of between 0% and +3.5%.

Overall, adding a rating factor to the FF6 baseline model yields limited, non-robust improvements that are merely confined to matched sustainability portfolios but not the broad standard test assets. The incremental contribution of a rating factor to FF6 baseline model is small and varies widely across providers and rating types.

5. Discussion and conclusion

Motivated by the methodological ambiguity in portfolio construction and the mixed empirical results concerning the relative performance of ESG portfolios and/or factors, this study uses a multi-specification approach to create portfolios and factors based on sustainability ratings. We first replicate the relevant analyses in three highly cited studies that examine portfolio and/or factor performance linked to (i) ESG rating and ESG rating uncertainty (Avramov et al., 2022), (ii) MSCI Environment ratings (Pástor et al., 2022), and (iii) sustainability rating disagreement (Gibson-Brandon et al., 2021). Employing the single-specification method documented in each study and our multi-specification approach, we find that rating-related portfolio/factor performance depends on specifications, factor models and sample periods. We obtain few robust positive outcomes, particularly in the out-of-sample period.

We further employ our multi-specification approach to create portfolios and factors for the analysis period July 2003-June 2024 using ENV, SOC, GOV, and ESG rating scores assigned by one of the five popular providers (Refinitiv, MSCI, RobecoSAM, Bloomberg, and Sustainalytics) or our self-constructed ENV/SOC/GOV/ESG Agreement/Disagreement scores. Covering 4,481 U.S. companies, we examine whether rating-sorted portfolios and rating factors generate positive excess returns/risk-adjusted returns and whether adding a rating factor can enhance the explanatory power of the Fama-French six-factor (FF6) model.

We do not find a consistent positive relationship between portfolios' rating ranks and portfolios' excess returns/risk-adjusted returns across different providers. The median excess returns/Sharpe ratios of sorted portfolios do not exhibit a clear and consistent upward trend moving from the lowest to the highest rating-ranked portfolios. Inconsistent patterns hold for rating-sorted decile portfolios and portfolios double-sorted by rating and size. Adding a rating factor to the base FF6 portfolio does not result in greater maximum ex-post Sharpe ratios. Using two alternative windows, the all-rating provider period (July 2017–June 2024) and the crisis period (January 2020–June 2024), all of our results remain qualitatively similar.

Factor spanning analysis indicates that rating factors are *not* entirely explained by the traditional FF6 factors, suggesting that rating factors *may* offer some unique information. However, the degree of this unique contribution varies widely across different providers and rating types. This inconsistency implies that while rating factors *may* provide additional insights, their impact is *not* uniformly recognized across different rating systems, limiting their overall usefulness in enhancing traditional asset pricing models.

Adding a rating factor to the FF6 model does *not* substantially improve the model's explanatory power. The adjusted R^2 values show negligible increases, whereas the number of Gibbons, Ross, and Shanken test rejections, indicating the model's ability to explain portfolio return variations, remains largely unchanged. For most rating categories and providers, adding a rating factor does not consistently and significantly reduce the degree of unexplained returns, suggesting that the FF6 model remains robust in explaining the cross-section of returns and that rating factors offer limited additional information beyond traditional factors.

Our analysis highlights that credible and robust assessments of sustainability rating performance require transparency about the underlying methodological choices. By reporting the entire spectrum of performance results across 512 factor specifications and all five major rating providers, we move beyond single-specification results and offer a more realistic understanding of what sustainability rating factors can and cannot deliver. As in Henriquez-Salman's (2025) study, we observe large differences in performance outcomes across construction choices. However, we enrich his findings by demonstrating that choices such as filtering criteria beyond sectors can significantly affect analysis outcomes.[47] For example, excluding versus including firms with negative earnings, while holding other decision points in Figure 1 constant, results in large differences in median monthly returns of -0.16%, -0.14%, and 0.21% for Bloomberg ESG, ENV, and GOV factors, respectively.

Our findings have practical implications for portfolio managers and institutional investors who often utilize sustainability ratings to make investment decisions. While our overall results show a non-positive (risk-adjusted) return performance, the performance outcome range across factor specifications is quite large. For example, over the horizon of 21 years, the worst performing REF, SUS, and AGR ENV factor specifications had returns of 1.7%, -58%, and -35.2%, while the best performing REF, SUS, and AGR ENV factor specifications delivered returns of 102%, 6.3%, and 26.4%, corresponding to respective ranges of 100.3%, 64.3%, and 61.6%. This emphasizes how sensitive sustainability rating-based investment strategies can be to arbitrary specifications. Without systematically testing alternative rating providers, rating categories, weighting schemes, and filter thresholds, practitioners may unknowingly select specifications that deliver extreme outcomes, leading to flawed investment decisions.

As our study focuses on U.S. firms, the findings may not fully generalize to markets with different cultural, economic, or regulatory environments. Moreover, our analysis does not consider transaction costs and other market frictions, which may affect the implementability of certain portfolio strategies. Additionally, our study reveals substantial variability across rating types and rating providers. The observed variability highlights the challenges in standardizing sustainability rating measurements, as different providers employ different methods and criteria in their assessment processes, leading to inconsistent implications for investment decisions. Though we analyze ratings issued by the five most popular agencies, including a broader range of providers may reveal more consistent patterns or confirm the inconsistencies across providers that we document.

[47] Henriquez-Salman (2025) does not employ any filter thresholds concerning firm size, book value, earnings, and exchanges for determining breakpoints.

References

Alves, Rómulo, Philipp Krüger, and Mathijs A. van Dijk. 2025. “Drawing Up the Bill: Are ESG Ratings Related to Stock Returns Around the World?” *Journal of Corporate Finance* 93: 102768. https://doi.org/10.1016/j.jcorpfin.2025.102768.

Atz, Ulrich, Tracy Van Holt, Zongyuan Zoe Liu, and Christopher C. Bruno. 2023. “Does Sustainability Generate Better Financial Performance? Review, Meta-Analysis, and Propositions.” *Journal of Sustainable Finance & Investment* 13 (1): 802–825. https://doi.org/10.1080/20430795.2022.2106934.

Avramov, Doron, Si Cheng, Abraham Lioui, and Andrea Tarelli. 2022. “Sustainable Investing with ESG Rating Uncertainty.” *Journal of Financial Economics* 145 (2): 642–664. https://doi.org/10.1016/j.jfineco.2021.09.009.

Bax, Karoline, Eleonora Broccardo, and Sandra Paterlini. 2024. “Environmental, Social, and Governance Factor and Financial Returns: What Is the Relationship? Investigating Environmental, Social, and Governance Factor Models.” *Current Opinion in Environmental Sustainability* 66: 101398. https://doi.org/10.1016/j.cosust.2023.101398.

Bell, Matthew, and Ben Taylor. 2024. “How Can Investors Balance Short-Term Demands with Long-Term Value?” EY Insights, Ernst & Young, December 10. https://www.ey.com/en_gl/insights/climate-change-sustainability-services/institutional-investor-survey.

Berg, Florian, Julian F. Kölbel, and Roberto Rigobon. 2022. “Aggregate Confusion: The Divergence of ESG Ratings.” *Review of Finance* 26 (6): 1315–1344. https://doi.org/10.1093/rof/rfac033.

Berg, Florian, Andrew W. Lo, Roberto Rigobon, Manish Singh, and Ruixun Zhang. 2024. “Quantifying the Returns of ESG Investing: An Empirical Analysis with Six ESG Metrics.” *Journal of Portfolio Management* 50(8): 216–238. https://doi.org/10.3905/jpm.2024.50.8.216.

Beyer, Victor, and Tobias Bauckloh. 2024. “Non-Standard Errors in Carbon Premia.” CFR Working Paper 24-06, Centre for Financial Research (CFR), University of Cologne. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4901081.

Busch, Timo, and Gunnar Friede. 2018. “The Robustness of the Corporate Social and Financial Performance Relation: A Second-Order Meta-Analysis.” *Corporate Social Responsibility and Environmental Management* 25 (4): 583–608. https://doi.org/10.1002/csr.1480.

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

Cakici, Nusret, Christian Fieberg, Gábor Neszveda, Vanja Piljak, and Adam Zaremba. 2025. “Lost in the Multiverse: Methodological Uncertainty in Studying Global Equity Returns.” *Critical Finance Review* (forthcoming). https://doi.org/10.2139/ssrn.5181455.

Ciciretti, Rocco, Ambrogio Dalò, and Lammertjan Dam. 2023. “The Contributions of Betas versus Characteristics to the ESG Premium.” *Journal of Empirical Finance* 71: 104–124. https://doi.org/10.1016/j.jempfin.2023.01.004.

Coqueret, Guillaume. 2022. *Perspectives in Sustainable Equity Investing*. Boca Raton, FL: CRC Press/Taylor & Francis.

Díaz, Violeta, Denada Ibrushi, and Jialin Zhao. 2021. “Reconsidering Systematic Factors during the COVID-19 Pandemic: The Rising Importance of ESG.” *Finance Research Letters* 38: 101870. https://doi.org/10.1016/j.frl.2020.101870.

Dobrick, Juris, Christian Klein, and Bernhard Zwergel. 2025. “ESG as Risk Factor.” *Journal of Asset Management* 26 (1): 44–70. https://doi.org/10.1057/s41260-024-00382-z.

Fama, Eugene F., and Kenneth R. French. 1993. "Common Risk Factors in the Returns on Stocks and Bonds." *Journal of Financial Economics* 33 (1): 3–56. https://doi.org/10.1016/0304-405X(93)90023-5.

Fama, Eugene F., and Kenneth R. French. 2007. "Disagreement, Tastes, and Asset Prices." *Journal of Financial Economics* 83 (3): 667–689. https://doi.org/10.1016/j.jfineco.2006.01.003.

Fama, Eugene F., and Kenneth R. French. 2008. "Dissecting Anomalies." *The Journal of Finance* 63 (4): 1653–1678. https://doi.org/10.1111/j.1540-6261.2008.01371.x

Fama, Eugene F., and Kenneth R. French. 2015. "A Five-Factor Asset Pricing Model." *Journal of Financial Economics* 116 (1): 1–22. https://doi.org/10.1016/j.jfineco.2014.10.010.

Fama, Eugene F., and Kenneth R. French. 2018. "Choosing Factors." *Journal of Financial Economics* 128 (2): 234–252. https://doi.org/10.1016/j.jfineco.2018.02.012.

Friede, Gunnar, Timo Busch, and Alexander Bassen. 2015. "ESG and Financial Performance: Aggregated Evidence from More Than 2000 Empirical Studies." *Journal of Sustainable Finance & Investment* 5 (4): 210–233. https://doi.org/10.1080/20430795.2015.1118917.

Gibson-Brandon, Rajna, Philipp Krueger, and Peter S. Schmidt. 2021. "ESG Rating Disagreement and Stock Returns." *Financial Analysts Journal* 77 (4): 104–127. https://doi.org/10.1080/0015198X.2021.1963186.

Global Sustainable Investment Alliance. 2023. *Global Sustainable Investment Review 2022*. London: Global Sustainable Investment Alliance. 1–47. Accessed June 5, 2025. https://www.gsi-alliance.org/wp-content/uploads/2023/12/GSIA-Report-2022.pdf.

Halbritter, Gerhard, and Gregor Dorfleitner. 2015. "The Wages of Social Responsibility — Where Are They? A Critical Review of ESG Investing." *Review of Financial Economics* 26 (1): 25–35. https://doi.org/10.1016/j.rfe.2015.03.004.

Hanauer, Matthias X. 2020. "A Comparison of Global Factor Models." SSRN Scholarly Paper 3546295, Social Science Research Network. https://doi.org/10.2139/ssrn.3546295.

Hasler, Mathias. 2023. "Looking Under the Hood of Data-Mining." *SSRN Scholarly Paper*, Revised September 24, 2023. https://doi.org/10.2139/ssrn.4279944.

Henriquez-Salman, Ricardo. 2025. "Methodological ESG Uncertainty in Portfolio Sorts." Research in International Business and Finance 80: 103132. https://doi.org/10.1016/j.ribaf.2025.103132.

Hou, Kewei, Chen Xue, and Lu Zhang. 2020. "Replicating Anomalies." *The Review of Financial Studies* 33 (5): 2019–2133. https://doi.org/10.1093/rfs/hhy131.

Hou, Kewei, Haitao Mo, Chen Xue, and Lu Zhang. 2021. "An Augmented q-Factor Model with Expected Growth." *Review of Finance* 25 (1): 1–41. https://doi.org/10.1093/rof/rfaa004.

Hübel, Benjamin, and Hendrik Scholz. 2020. "Integrating Sustainability Risks in Asset Management: The Role of ESG Exposures and ESG Ratings." *Journal of Asset Management* 21 (1): 52–69. https://doi.org/10.1057/s41260-019-00139-z.

Ince, Ozgur S., and R. Burt Porter. 2006. "Individual Equity Return Data from Thomson Datastream: Handle with Care!" *Journal of Financial Research* 29 (4): 463–479. https://doi.org/10.1111/j.1475-6803.2006.00189.x.

Index Industry Association. 2023. "IIA's Third Annual ESG Survey of Global Asset Managers Reveals That Widening Factors, Expanding Asset Classes, and Emerging Tech Are Driving a Maturing ESG Landscape." Press release, June 27. https://www.indexindustry.org/iias-third-annual-esg-survey-of-global-asset-managers-reveals-that-widening-factors-expanding-asset-classes-and-emerging-tech-are-driving-a-maturing-esg-landscape-2/.

Jin, Ick. 2018. "Is ESG a Systematic Risk Factor for U.S. Equity Mutual Funds?" Journal of *Sustainable Finance & Investment* 8 (1): 72–93. https://doi.org/10.1080/20430795.2017.1395251.

Khan, Mozaffar. 2019. "Corporate Governance, ESG, and Stock Returns around the World." *Financial Analysts Journal* 75 (4): 103–123. https://doi.org/10.1080/0015198X.2019.1654299.

Kumar, Sumit. 2023. "Exploratory Review of ESG Factor Attribution to the Portfolio Return in Fama-French Factor Model Framework." *Academy of Marketing Studies Journal* 27 (Special Issue 3): 1–20.

Landis, Conrad, and Spyros Skouras. 2021. "Guidelines for Asset Pricing Research Using International Equity Data from Thomson Reuters Datastream." *Journal of Banking & Finance* 130: 106128. https://doi.org/10.1016/j.jbankfin.2021.106128.

Liang, Hao, and Luc Renneboog. 2021. "Corporate Social Responsibility and Sustainable Finance." In *Oxford Research Encyclopedia of Economics and Finance*. Oxford: Oxford University Press. https://doi.org/10.1093/acrefore/9780190625979.013.592.

Lioui, Abraham, and Andrea Tarelli. 2022. "Chasing the ESG Factor." *Journal of Banking & Finance* 139: 106498. https://doi.org/10.1016/j.jbankfin.2022.106498.

Madhavan, Ananth, Aleksander Sobczyk, and Andrew Ang. 2021. "Toward ESG Alpha: Analyzing ESG Exposures through a Factor Lens." *Financial Analysts Journal* 77 (1): 69–88. https://doi.org/10.1080/0015198X.2020.1816366.

Maiti, Moinak. 2021. "Is ESG the Succeeding Risk Factor?" *Journal of Sustainable Finance & Investment* 11 (3): 199–213. https://doi.org/10.1080/20430795.2020.1723380.

Margolis, Joshua D., Hillary Anger Elfenbein, and James P. Walsh. 2009. “Does It Pay to Be Good … And Does It Matter? A Meta-Analysis of the Relationship between Corporate Social and Financial Performance.” SSRN Scholarly Paper no. 1866371. https://doi.org/10.2139/ssrn.1866371.

Naffa, Helena, and Máté Fain. 2020. “Performance Measurement of ESG-Themed Megatrend Investments in Global Equity Markets Using Pure Factor Portfolios Methodology.” *PLOS ONE* 15 (12): e0244225. https://doi.org/10.1371/journal.pone.0244225.

Naffa, Helena, and Máté Fain. 2022. “A Factor Approach to the Performance of ESG Leaders and Laggards.” *Finance Research Letters* 44: 102073. https://doi.org/10.1016/j.frl.2021.102073.

Nsibande, Luyanda M. Q., and Avani Sebastian. 2023. “Is the Environmental, Social and Corporate Governance Score the Missing Factor in the Fama-French Five-Factor Model?” *South African Journal of Economic and Management Sciences* 26 (1): a4835. https://doi.org/10.4102/sajems.v26i1.4835.

Pástor, Ľuboš, Robert F. Stambaugh, and Lucian A. Taylor. 2021. “Sustainable Investing in Equilibrium.” *Journal of Financial Economics* 142 (2): 550–571. https://doi.org/10.1016/j.jfineco.2020.12.011.

Pástor, Ľuboš, Robert F. Stambaugh, and Lucian A. Taylor. 2022. “Dissecting Green Returns.“ *Journal of Financial Economics* 146 (2): 403–424. https://doi.org/10.1016/j.jfineco.2022.07.007.

Pedersen, Lasse Heje, Shaun Fitzgibbons, and Łukasz Pomorski. 2021. “Responsible Investing: The ESG-Efficient Frontier.” *Journal of Financial Economics* 142 (2): 572–597. https://doi.org/10.1016/j.jfineco.2020.11.001.

Pollard, Julia L., Matthew W. Sherwood, and Ryan G. Klobus. 2018. “Establishing ESG as Risk Premia.” *Journal of Investment Management* 16 (1): 32–43.

Walter, Dominik, Rüdiger Weber, and Patrick Weiss. 2024. “Methodological Uncertainty in Portfolio Sorts.” SSRN Scholarly Paper no. 4164117. https://doi.org/10.2139/ssrn.4164117.

Wong, Christina, Aiste Brackley, and Erika Petroy. 2019. “Rate the Raters 2019: Expert Views on ESG Ratings.” New York: Sustainability, available at https://www.sustainability.com/globalassets/sustainability.com/thinking/pdfs/sa-ratetheraters-2019-1.pdf.

Xiao, Yuchao, Robert Faff, Philip Gharghori, and Darren Lee. 2013. “An Empirical Study of the World Price of Sustainability.” *Journal of Business Ethics* 114 (2): 297–310. https://doi.org/10.1007/s10551-012-1342-2.

Table 1. Rating factor spanning

Provider, analysis period	**Rating**	**RM**	**SMB5**	**HML**	**RMW**	**CMA**	**MOM**	**R^2**
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
REF July 2003- June 2024, Obs: 252	ESG	0.018 (23%)	–0.123 (94%)	0.095 (65%)	0.173 (75%)	0.250 (100%)	–0.019 (25%)	0.302
	ENV	0.086 (89%)	–0.162 (93%)	0.217 (95%)	0.301 (100%)	0.189 (90%)	–0.014 (12%)	0.421
	SOC	0.020 (33%)	–0.102 (79%)	0.029 (8%)	0.046 (26%)	0.117 (58%)	–0.027 (34%)	0.129
	GOV	0.006 (2%)	–0.054 (31%)	0.126 (76%)	0.232 (98%)	0.149 (76%)	0.041 (54%)	0.341
MSCI July 2008– June 2024, Obs: 192	ESG	–0.011 (13%)	–0.081 (53%)	0.011 (7%)	0.085 (44%)	–0.020 (4%)	0.029 (29%)	0.133
	ENV	–0.024 (12%)	–0.214 (100%)	–0.159 (82%)	–0.236 (89%)	0.012 (7%)	0.018 (27%)	0.343
	SOC	–0.001 (26%)	–0.011 (1%)	–0.062 (49%)	0.063 (29%)	0.042 (0%)	–0.007 (28%)	0.111
	GOV	–0.021 (11%)	0.068 (38%)	0.115 (79%)	0.171 (95%)	–0.154 (91%)	0.029 (30%)	0.125
BB July 2016-June 2024, Obs: 96	ESG	–0.045 (30%)	–0.145 (51%)	–0.023 (23%)	0.065 (5%)	0.108 (17%)	–0.051 (12%)	0.191
	ENV	0.002 (9%)	–0.143 (48%)	0.041 (61%)	0.196 (65%)	0.171 (48%)	–0.168 (100%)	0.363
	SOC	0.041 (25%)	0.028 (0%)	0.195 (84%)	0.167 (45%)	0.046 (10%)	–0.130 (75%)	0.432
	GOV	–0.076 (74%)	–0.033 (4%)	0.106 (56%)	0.244 (85%)	0.282 (100%)	–0.001 (8%)	0.586
ROB July 2017- June 2024, Obs: 84	ESG	–0.026 (12%)	–0.204 (99%)	0.066 (28%)	0.027 (5%)	0.093 (31%)	–0.111 (70%)	0.309
	ENV	–0.033 (9%)	–0.299 (100%)	0.045 (2%)	–0.105 (28%)	0.091 (19%)	–0.132 (74%)	0.301
	SOC	–0.028 (8%)	–0.201 (78%)	0.128 (49%)	–0.037 (0%)	0.092 (32%)	–0.112 (60%)	0.282
	GOV	–0.023 (3%)	–0.081 (16%)	0.033 (5%)	0.069 (28%)	0.159 (52%)	–0.087 (54%)	0.258
SUS July 2010-June 2024, Obs: 168	ESG	0.045 (21%)	–0.176 (77%)	–0.047 (4%)	0.015 (8%)	–0.013 (0%)	–0.140 (98%)	0.136
	ENV	0.025 (0%)	–0.181 (92%)	–0.106 (38%)	–0.066 (12%)	–0.005 (0%)	–0.074 (34%)	0.121
	SOC	0.029 (22%)	–0.095 (12%)	0.069 (2%)	0.068 (10%)	–0.147 (34%)	–0.126 (100%)	0.134
	GOV	–0.041 (35%)	–0.028 (2%)	–0.101 (50%)	0.251 (83%)	0.144 (48%)	–0.078 (55%)	0.198

Table 1. Rating factor spanning (continued)

Provider, analysis period	Rating	RM	SMB5	HML	RMW	CMA	MOM	R²
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Rating Agreement July 2010 – June 2024, Obs: 168	ESG	–0.036 (30%)	–0.134 (89%)	0.018 (12%)	0.172 (71%)	0.112 (38%)	–0.052 (33%)	0.252
	ENV	–0.022 (23%)	–0.211 (100%)	–0.061 (38%)	0.092 (36%)	0.229 (94%)	–0.074 (59%)	0.308
	SOC	–0.037 (34%)	–0.083 (39%)	0.106 (62%)	0.152 (73%)	0.076 (17%)	–0.077 (67%)	0.26
	GOV	–0.041 (42%)	–0.010 (2%)	0.135 (76%)	0.269 (100%)	0.139 (68%)	0.020 (3%)	0.373
Rating Disagreement July 2010 – June 2024, Obs: 168	ESG	0.023 (10%)	–0.041 (0%)	–0.006 (0%)	–0.048 (1%)	0.054 (9%)	–0.018 (1%)	0.034
	ENV	–0.004 (9%)	–0.015 (11%)	–0.123 (85%)	–0.098 (54%)	0.037 (4%)	0.042 (19%)	0.185
	SOC	–0.012 (0%)	–0.118 (96%)	–0.104 (71%)	–0.016 (5%)	0.073 (23%)	–0.032 (4%)	0.156
	GOV	–0.042 (49%)	–0.042 (14%)	–0.038 (16%)	–0.031 (21%)	–0.016 (0%)	0.001 (0%)	0.106

Explanation: This table presents the median coefficient estimates derived from estimating regressions of a rating factor (ENV, SOC, GOV, ESG) on Fama-French's (FF) six factors (FF6) namely excess market return (RM), Small Minus Big (SMB5 in the FF five-factor model), High Minus Low (HML), Robust Minus Weak (RMW), Conservative Minus Aggressive (CMA), and Momentum (MOM). We create 512 factor specifications for each provider and each rating type. For each model, we report the median coefficient estimate across 512 factor specifications and indicate in parentheses the percentage of specifications for which the concerned coefficient estimate is significant at the 5% level.

We normalize all ratings before constructing rating factors. Each factor is computed as the average return difference between the two High-Rating (Big-High and Small-High) portfolios and the two Low-Rating (Big-Low and Small-Low) portfolios, as outlined in section 4.3.1. We conduct separate analyses and report the corresponding results for rating factors specific to Refinitiv (REF), MSCI, Bloomberg (BB), RobecoSAM (ROB), Sustainalytics (SUS), as well as our self-constructed rating Agreement and Disagreement scores.

Interpretation: Factors from the FF6 model partially explain each rating-sorted factor. We observe a wide variability across providers and rating types in terms of the median R^2 and the significance of the FF6 factor loadings which range between 0% and 100%. As an example, refer to CMA estimates in column 7. for BB GOV factor, 100% of CMA estimates are significant while for the DIS GOV factor, none of the CMA estimates are significant.

Table 2. GRS test: Incremental contribution of a rating factor to the Fama-French six-factor (FF6) model

	Sustainability Rating Test Sets (FF6 with A Rating Factor – FF6)							Standard Test Sets (FF6 with A Rating Factor – FF6)						
	%Δ (Δ) p(GRS-Test) <5%	Δ $\overline{R^2}$	Δ $\overline{R_H^2 - R_L^2}$	%Δ (Δ) $p(\alpha) < 1\%$	%Δ (Δ) $p(\alpha) < 5\%$	Δ $\overline{\lvert\alpha\rvert}$	Δ $\overline{\alpha_H - \alpha_L}$	%Δ (Δ) p(GRS-Test) <5%	Δ $\overline{R^2}$	Δ $\overline{R^2{}_H - R^2{}_L}$	%Δ (Δ) $p(\alpha) < 1\%$	%Δ (Δ) $p(\alpha) < 5\%$	Δ $\overline{\lvert\alpha\rvert}$	Δ $\overline{\alpha_H - \alpha_L}$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
Panel A: Refinitiv (July 2003 – June 2024)														
ESG	–4.5% (–46)	0.02	0.00	–0.7% (–95)	–0.7% (–99)	0.01	0.08	–0.5% (–1)	0.00	–0.01	–0.2% (–5)	–0.5% (–10)	0.00	–0.01
ENV	–31.2% (–319)	0.03	–0.03	–2.1% (–276)	–6% (–798)	–0.02	–0.05	2% (4)	0.00	–0.01	0.3% (6)	0.2% (5)	0.00	0.01
SOC	4.2% (43)	0.02	0.00	0% (–2)	0.7% (93)	0.02	0.12	0% (0)	0.00	–0.01	–0.1% (–3)	–0.4% (–8)	0.00	–0.01
GOV	–8.8% (–90)	0.02	–0.01	–1.6% (–211)	–3.1% (–415)	0.01	0.02	0.5% (1)	0.00	–0.01	–0.1% (–2)	–0.4% (–8)	0.00	0.00
Panel B: MSCI (July 2008 – June 2024)														
ESG	4.7% (48)	0.01	–0.01	0.8% (111)	0.1% (11)	–0.01	–0.02	–1.5% (–3)	0.00	0.00	–0.1% (–3)	–0.2% (–5)	0.00	0.00
ENV	–2% (–20)	0.01	0.00	0.1% (12)	0.2% (25)	0.00	0.01	–1.5% (–3)	0.00	0.00	0% (–1)	–0.1% (–3)	0.00	0.00
SOC	–1% (–10)	0.01	0.00	–0.5% (–70)	–0.3% (–44)	0.00	0.00	–1.5% (–3)	0.00	0.00	–0.2% (–5)	–0.4% (–9)	0.00	–0.01
GOV	–2% (–20)	0.02	0.00	0.7% (94)	0.5% (67)	0.00	0.03	–1% (–2)	0.00	0.00	0% (1)	0.2% (5)	0.00	0.00
Panel C: Bloomberg (July 2016 – June 2024)														
ESG	1% (10)	0.01	–0.01	0.3% (41)	0.4% (53)	0.00	–0.01	5% (10)	0.00	0.00	0.6% (13)	1.7% (36)	0.01	0.04
ENV	–0.5% (–5)	0.01	–0.01	0.2% (21)	0.4% (49)	0.00	–0.01	2% (4)	0.00	0.00	0.1% (2)	0.2% (4)	0.01	0.02
SOC	0.6% (6)	0.00	–0.01	0% (3)	0% (4)	0.00	–0.01	2% (4)	0.00	0.00	0.1% (3)	0.2% (4)	0.00	0.01
GOV	–1.2% (–12)	0.01	0.00	0.1% (12)	–0.1% (–10)	0.00	–0.02	2% (4)	0.00	0.00	–0.1% (–2)	0% (1)	0.01	0.02
Panel D: RobecoSAM (July 2017 – June 2024)														
ESG	0.1% (1)	0.00	–0.02	0.1% (12)	–0.1% (–15)	0.00	–0.01	0% (0)	0.00	0.00	0% (–1)	0.1% (3)	0.00	0.00
ENV	–0.1% (–1)	0.01	0.00	0.7% (93)	0.1% (11)	–0.01	0.00	0% (0)	0.01	0.00	0% (–1)	0% (1)	0.00	–0.01
SOC	0.7% (7)	0.01	–0.01	0.5% (68)	–0.3% (–39)	0.00	–0.02	–0.5% (–1)	0.01	0.00	–0.2% (–4)	–0.2% (–4)	0.00	–0.01
GOV	–2.1% (–21)	0.00	–0.01	0% (1)	–0.1% (–10)	–0.01	–0.04	0.5% (1)	0.00	0.00	0% (–1)	0% (–1)	0.00	0.00
Panel E: Sustainalytics (July 2010 – June 2024)														
ESG	–3.4% (–35)	0.01	0.00	0.1% (14)	0.5% (67)	0.01	0.04	1% (2)	0.00	0.00	0.1% (3)	0.6% (12)	0.01	0.01
ENV	–7.4% (–76)	0.00	–0.03	–0.5% (–63)	1.9% (247)	0.01	0.02	2.5% (5)	0.00	0.00	0.6% (12)	1% (20)	0.00	0.01
SOC	–8.4% (–86)	0.01	–0.01	–1.7% (–231)	–2% (–260)	0.00	0.03	3.5% (7)	0.00	0.00	0.5% (10)	1.4% (30)	0.01	0.02
GOV	–11.9% (–122)	0.00	0.01	–0.9% (–125)	–0.5% (–68)	0.00	0.02	0% (0)	0.00	0.00	–0.1% (–3)	0% (0)	0.00	–0.01

Table 2. GRS test: Incremental contribution of a rating factor to the Fama-French six-factor (FF6) model (continued)

	Sustainability Rating Test Sets (FF6 with A Rating Factor – FF6)							Standard Test Sets (FF6 with A Rating Factor – FF6)						
	%Δ (Δ)	Δ	Δ	%Δ (Δ)	%Δ (Δ)	Δ	Δ	Δ	Δ	Δ	%Δ (Δ)	%Δ (Δ)	Δ	Δ
	p(GRS-Test) <5%	$\overline{R^2}$	$\overline{{R^2}_H - {R^2}_L}$	$p(\alpha) < 1\%$	$p(\alpha) < 5\%$	$\overline{\lvert\alpha\rvert}$	$\overline{\alpha_H - \alpha_L}$	p(GRS-Test) <5%	$\overline{R^2}$	$\overline{{R^2}_H - {R^2}_L}$	$p(\alpha) < 1\%$	$p(\alpha) < 5\%$	$\overline{\lvert\alpha\rvert}$	$\overline{\alpha_H - \alpha_L}$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
Panel F: Rating Agreement (July 2010 – June 2024)														
ESG	–2.1% (–22)	0.01	–0.01	0% (–3)	–0.3% (–43)	0.00	0.01	4.5% (9)	0.00	0.00	0.8% (17)	1.4% (29)	0.00	0.01
ENV	0.5% (5)	0.01	0.01	0% (3)	–0.2% (–29)	0.01	0.01	2% (4)	0.00	0.00	0.3% (7)	0.5% (11)	0.00	0.00
SOC	2.6% (27)	0.01	0.00	0.1% (14)	0.3% (36)	0.01	0.02	5% (10)	0.00	0.00	0.8% (16)	1.1% (23)	0.00	0.01
GOV	–0.4% (–4)	0.01	0.00	–0.3% (–43)	–1.1% (–152)	0.01	0.00	2.5% (5)	0.00	0.00	0.3% (7)	0.1% (3)	0.00	–0.01
Panel G: Rating Disagreement (July 2010 – June 2024)														
ESG	0.5% (5)	0.01	0.00	0.2% (24)	0.4% (55)	0.00	0.02	0.5% (1)	0.00	0.00	0% (1)	0% (0)	0.00	–0.01
ENV	–0.2% (–2)	0.01	0.00	0.1% (14)	0.7% (96)	0.01	0.04	1.5% (3)	0.00	0.00	0.3% (6)	0.2% (5)	0.00	0.00
SOC	4.9% (50)	0.01	0.00	0.3% (43)	0.6% (82)	0.01	0.04	0% (0)	0.00	0.00	0% (–1)	0.1% (2)	0.00	0.00
GOV	–2.9% (–30)	0.01	0.00	–0.2% (–29)	–0.9% (–117)	0.00	0.01	0.5% (1)	0.00	0.00	0% (1)	–0.1% (–2)	0.00	0.00

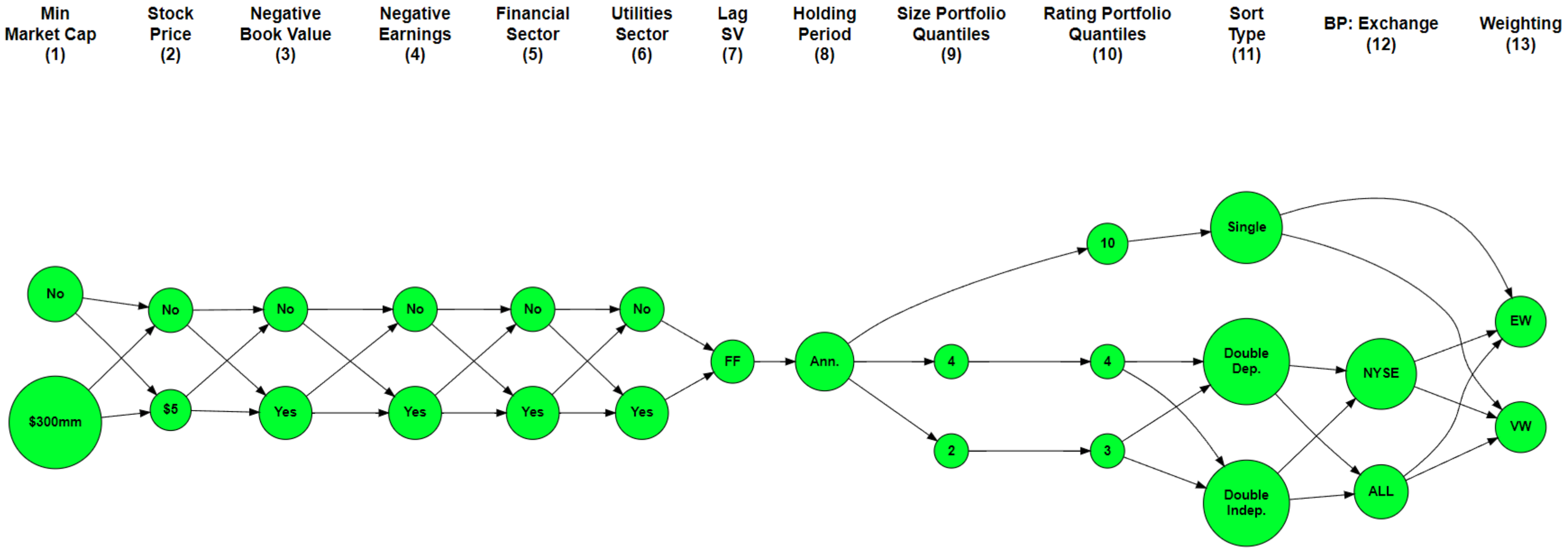
Explanation: This table reports the incremental contribution of adding a sustainability rating factor to the baseline FF6 model. Reported statistics capture differences (Δ) between the extended model and the baseline FF6 model (FF6+a Rating Factor – FF6). The *p(GRS-Test)* tests the joint hypothesis that all portfolio alphas are equal to zero. %Δ (Δ)$p(GRS\text{-}Test) < 5\%$ (columns 2 and 9) show the difference in percentage (number) of sets for which the hypothesis is rejected (the *p*-values from the GRS are lower than 5% for each test set). Δ $\overline{R^2}$ (columns 3 and 10) is the difference in average adjusted R^2 across specifications for each test set. Δ $\overline{R^2_H - R^2_L}$ (columns 4 and 11) is the difference in average difference between the highest and lowest R^2 across specifications for each test set. %Δ (Δ)$p(\alpha) < 1\%$ (columns 5 and 12) and %Δ (Δ)$p(\alpha) < 5\%$ (columns 6 and 13) are the difference in percentage (number) of portfolios with alpha being statistically and significantly different from zero at the 1% and 5% level, respectively. Δ$\overline{|\alpha|}$ (columns 7 and 14) is the difference in average absolute alpha across specifications for each test set, and Δ$\overline{\alpha_H - \alpha_L}$ (columns 8 and 15) is the difference in average difference between the highest and lowest alpha across specifications for each test set. Negative values for Δ$p(GRS\text{-}Test) < 5\%$ and Δ$p(\alpha) < 5\%/1\%$ indicate fewer rejections of the GRS-Test when a rating factor is added to the baseline model. Positive values for Δ $\overline{R^2}$ indicate higher explanatory power, while negative values for Δ$\overline{R^2_H - R^2_L}$, Δ$\overline{|\alpha|}$, and Δ$\overline{\alpha_H - \alpha_L}$ indicate reductions in cross-sectional mispricing.

The left block (Sustainability Rating Test Sets, columns 2-8) evaluates each provider–rating factor specification (512 specifications per rating type) against 1024 sustainability rating portfolio test sets which include 512 single-sorted decile portfolios (10×1) and 512 double-sorted quartile portfolios (4×4), totaling 13,312 portfolios. Each portfolio corresponds to the respective rating type (e.g., ESG factors are tested against ESG portfolios, and so on), and a factor specification is always paired with the corresponding portfolio specification (e.g., REF ESG Factor specification 5 of 512 is evaluated against REF ESG Portfolio specification 5 of 512). To ensure comparability, results are normalized by dividing by 7, since each rating factor is tested against portfolios from five providers and our self-calculated rating Agreement and Disagreement scores.

The right block (Standard Test Sets, columns 9-15) evaluates each of the 512 factor specifications against the same 200 standard test portfolio sets which consist of 7 double-sorted 5x5 sets from French's data library (https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html) and 193 sets from Global-Q (https://global-q.org/testingportfolios.html), totaling 2073 portfolios. Results are normalized by dividing by the number of specifications (512) to make them comparable to the FF6 baseline.

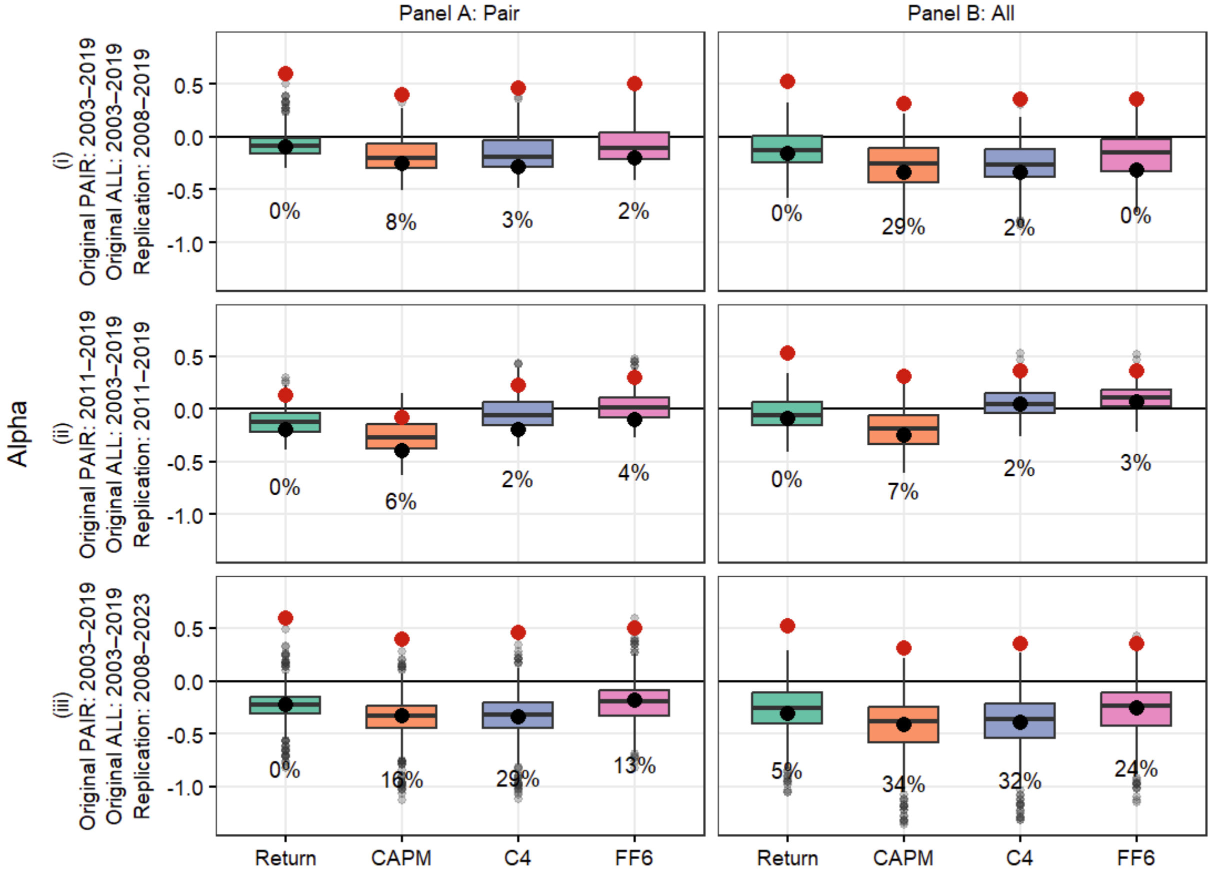
Interpretation: Adding a rating factor to the FF6 baseline model yields marginal improvements that are mostly confined to the matched sustainability rating portfolio sets. For the standard portfolio sets, there is no economically meaningful incremental fit or reduction in pricing errors.

Figure 1. Multi-specification portfolio construction procedure



Explanation: This flowchart depicts all decision points (nodes) that we consider in constructing our rating-sorted portfolios and subsequently, our rating-sorted factors (ENV, SOC, GOV, ESG score/rating Agreement/Disagreement). In each step (a node) between (1) and (6), we consider two options. First, whether to exclude small-cap firms with market capitalization of less than US$ 300 million or to include all firms (1). Second, whether to exclude penny stocks with price of less than US$ 5 or to include all (2). Next, whether to exclude firms with negative book value (3), negative earnings (4), or firms in the financial (5) or utilities (6) sector. We also consider how long to lag the sorting variable (SV) (node 7). For each rating provider, we use the rating scores in December of year $t_{(-1)}$ to construct the portfolio in June of year t_0 for the period from July of year t_0 to June of year t_1 (similar to Fama and French (FF), 1993). We rebalance our rating portfolio annually (Ann.) considering the annual sustainability rating data availability (node 8). We then consider the number of quantiles to construct a portfolio. If we single-sort, we use deciles based on rating scores (10x1, nodes 10-11). If we double-sort, we either utilize four quantiles for ratings and four quantiles for size (4x4, nodes 9-10) or two 50/50 size portfolios and three rating portfolios formed at the usual 30 (low)/40 (medium)/30 (high) split for the ratings as in Fama and French (1993) (nodes 9-10). The former (latter) results in 16 4x4 (six 2x3) portfolios (nodes 9-10). Double-sorted portfolios are either dependently sorted or independently sorted (node 11). In a dependent sort, we first sort by size, and within each size portfolio, sort by the rating. In independent double-sorts, we form the intersection of the size and rating portfolios (Small-High, Small-Low, Big-High, Big-Low). For double-sorted portfolios, the breaking points (BP) for the size portfolios can be based on either only stocks traded on the New York Stock Exchange (NYSE) or on ALL stock exchanges (node 12). The final decision is whether to value weight (VW) or equally weight (EW) stocks (node 13). Overall, we have 128 specifications (2^7=128) of decile portfolios and 512 specifications (2^9=512) of two-way sorted 4x4 and 2x3 portfolios.

Figure 2. Avramov et al. (2022) Replication: *Low ESG Uncertainty, Low ESG Rating* minus *Low ESG Uncertainty, High ESG Rating* return (*Low LMH-R* return)



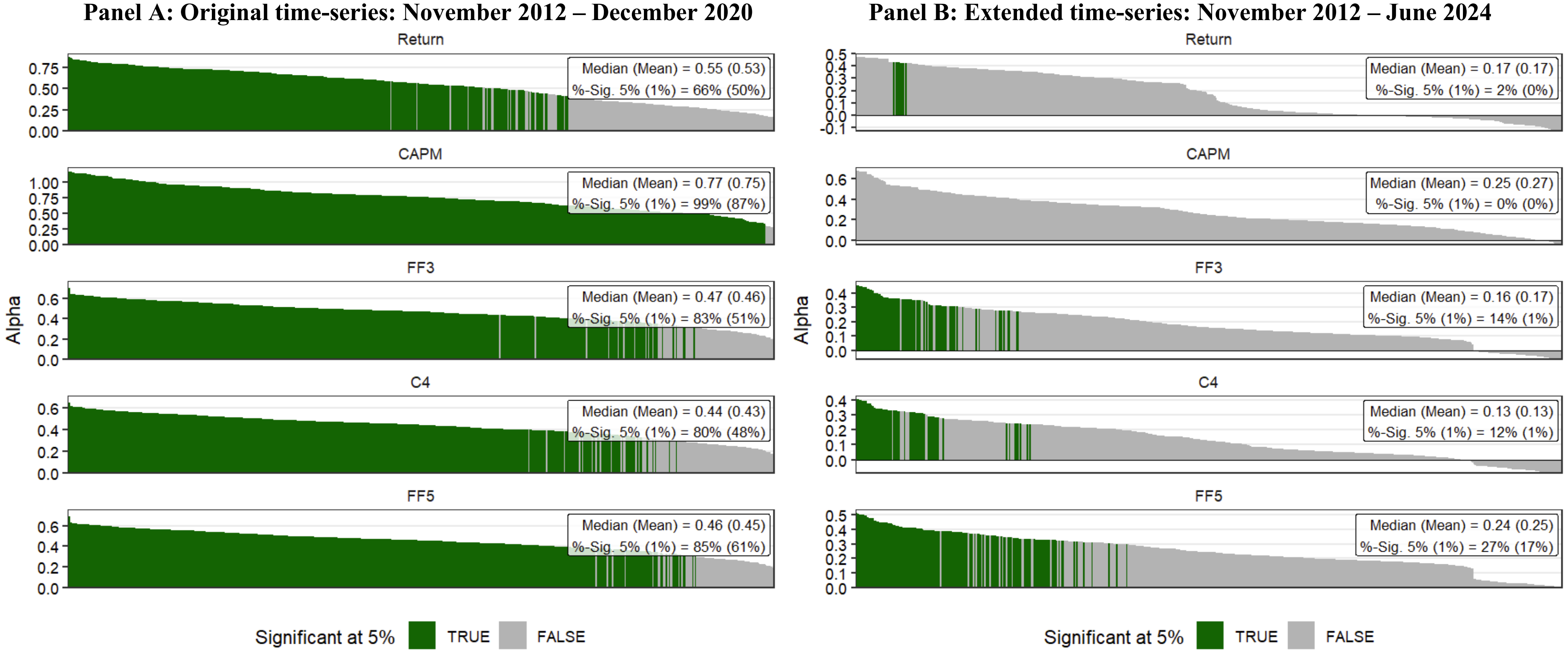
Explanation: This figure summarizes the long-short *Low LMH-R* (*Low ESG Uncertainty, Low ESG Rating* minus *Low ESG Uncertainty, High ESG Rating*) alphas reported in Avramov et al. (2022), and replicated alphas obtained by using their single specification method and our 512-specification method. For each panel, we plot boxplots of the distribution of 512-specification return and alpha estimates derived from four models: Return (model with constant), CAPM (CAPM), Fama-French-Carhart four-factor (C4) model, and Fama-French six-factor (FF6) model. The number printed below each box is the share of our 5%-significant alpha estimates for that specification set. The black dots are our estimates using Avramov et al.'s (2022) single specification and our samples. The red dots mark the point estimates reported by the authors using their single-specification approach and their samples. The red dots correspond to the following tables in Avramov et al. (2022):

- Table 2 (2003–2019): Return (Panel A) and CAPM-adjusted return (Panel B).
- Table 5 (2011–2019): Return (Panel A) and CAPM-adjusted return (Panel B).
- Table B.4 (2003–2019): Carhart-4 (Panel A) and FF6 (Panel B) alphas.
- Table B.5 (2011–2019): Carhart-4 (Panel A) and FF6 (Panel B) alphas.
- Table B.7 (2003–2019, "ALL" variants): Return (Panel A), CAPM (Panel B), Carhart-4 (Panel C), and FF6 (Panel D).

Panel A PAIR (main measure) and Panel B ALL (alternative measure for robustness analysis) refer to how ESG rating and ESG rating-uncertainty are constructed by the authors. Three parts in Panel A correspond to three sample periods: (i) original study period 2003–2019 versus our replication period 2008–2019; (ii) a directly comparable window, original study 2011–2019 versus our replication 2011–2019; and (iii) original study 2003–2019 versus our extended period 2008–2023. Parts (i) and (iii) in Panel B use the same respective sample periods as parts (i) and (iii) in Panel A. Part (ii) of Panel B (ALL) utilizes the original study period 2003–2019 versus our replication period 2011–2019 as the authors do not create a subperiod sample for their robustness measure ALL.

Interpretation: In contrast to the positive *Low LMH-R* estimates reported in Avramov et al. (2022), our 512-specification replications for the PAIR and ALL measures show mostly negative and rarely significant *Low LMH-R* returns/alphas in all windows. Alpha estimates disperse widely, indicating their strong sensitivity to specifications and sample periods.

Figure 3. Pástor et al. (2022) Table 3 Replication – Green minus Brown (GMB)'s return and alpha distribution using our multi-specification method



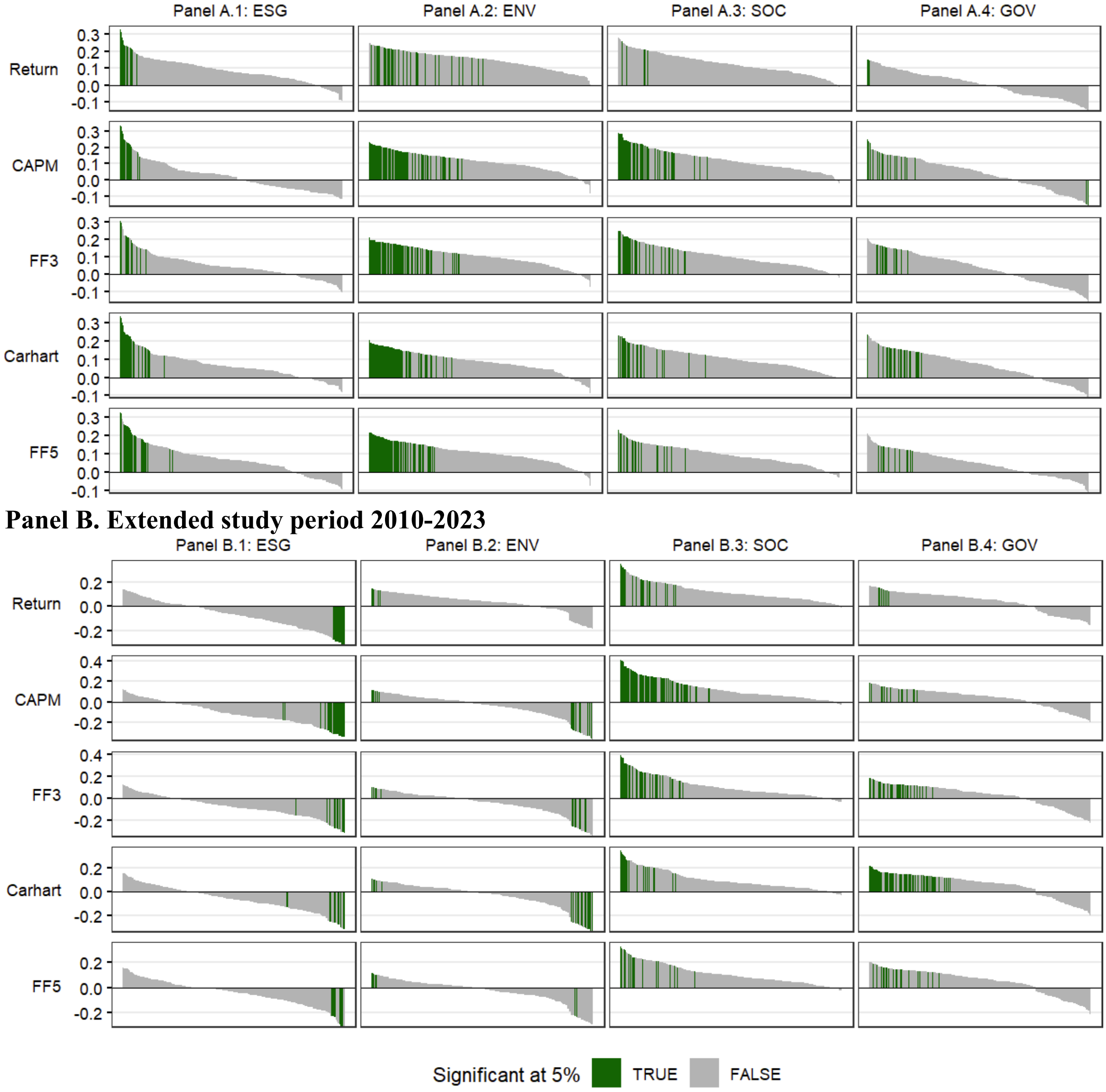
Panel A: Original time-series: November 2012 – December 2020

Panel B: Extended time-series: November 2012 – June 2024

Explanation: This appendix shows the distribution of 512 alpha estimates for each factor model corresponding to Pástor et al. (2022) Table 3, columns 1 (Return, model with constant), 2 (CAPM), 3 (Fama-French three-factor (FF3) model,), 4 (Carhart four-factor (C4) model), and 6 (Fama-French five-factor (FF5) model), with the dependent variable being the respective GMB spread specification. Each bar plot is sorted by its alpha values from the highest to the lowest. A significant alpha (at the 5% level) is denoted in green, and a non-significant one is in grey. The median and mean alphas and the proportions of alphas that are significant at the 5% and 1% level are included in the top right corner of each plot. We estimate each model, with the Newey–West standard errors, for the original study period ending in December 2020 (Panel A) and the extended period ending in June 2024 (Panel B).

Interpretation: During the original study period, November 2012 - December 2020, significant alpha estimates are dominant; however, *not* all alpha estimates are significant. During our extended period ending in June 2024, the mean/median alphas and the proportions of significant alpha estimates across models declined substantially, suggesting that alpha estimates are highly dependent on specifications, factor models, and sample periods.

Figure 4. Gibson-Brandon et al. (2021) Table 5 Replication – Alpha estimate distribution using our multi-specification approach



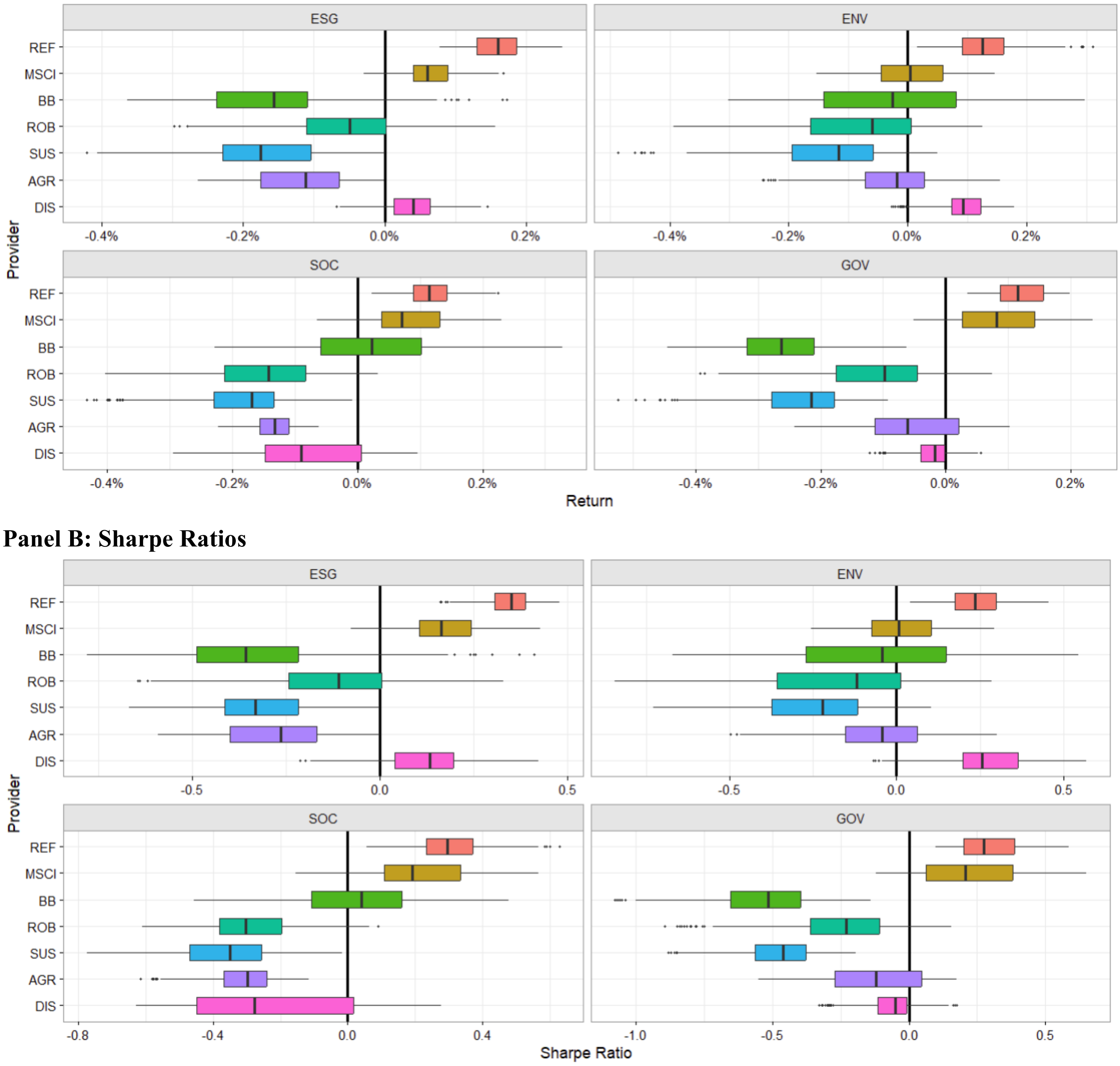
Panel A. Original study period 2010-2017

Panel B. Extended study period 2010-2023

Explanation: This figure features the distribution of alpha estimates derived from our multi-specification analysis of return differences between high and low industry-adjusted rating disagreement portfolios. Each bar plot is sorted by its alpha values from the highest to the lowest. For each rating category (ESG, ENV, SOC, GOV), we estimate a model with Return (model with constant), CAPM, Fama-French three-factor (FF3), Carhart four-factor (C4), and Fama-French five-factor (FF5) models. We form portfolios during the original study period of January 2010-December 2017 (Panel A) and an extended period of January 2010-December 2023 (Panel B). A significant alpha (at the 5% level) is denoted in green and a non-significant one is in grey. All models are estimated using the Newey–West standard errors.

Interpretation: Our multi-specification results show that alpha estimates are highly dependent on the specifications used in portfolio construction and sample periods. Most alpha estimates, particularly in the extended period 2010-2013, are insignificant.

Figure 5. Rating factors: returns and Sharpe ratios



Panel A: Factor Returns

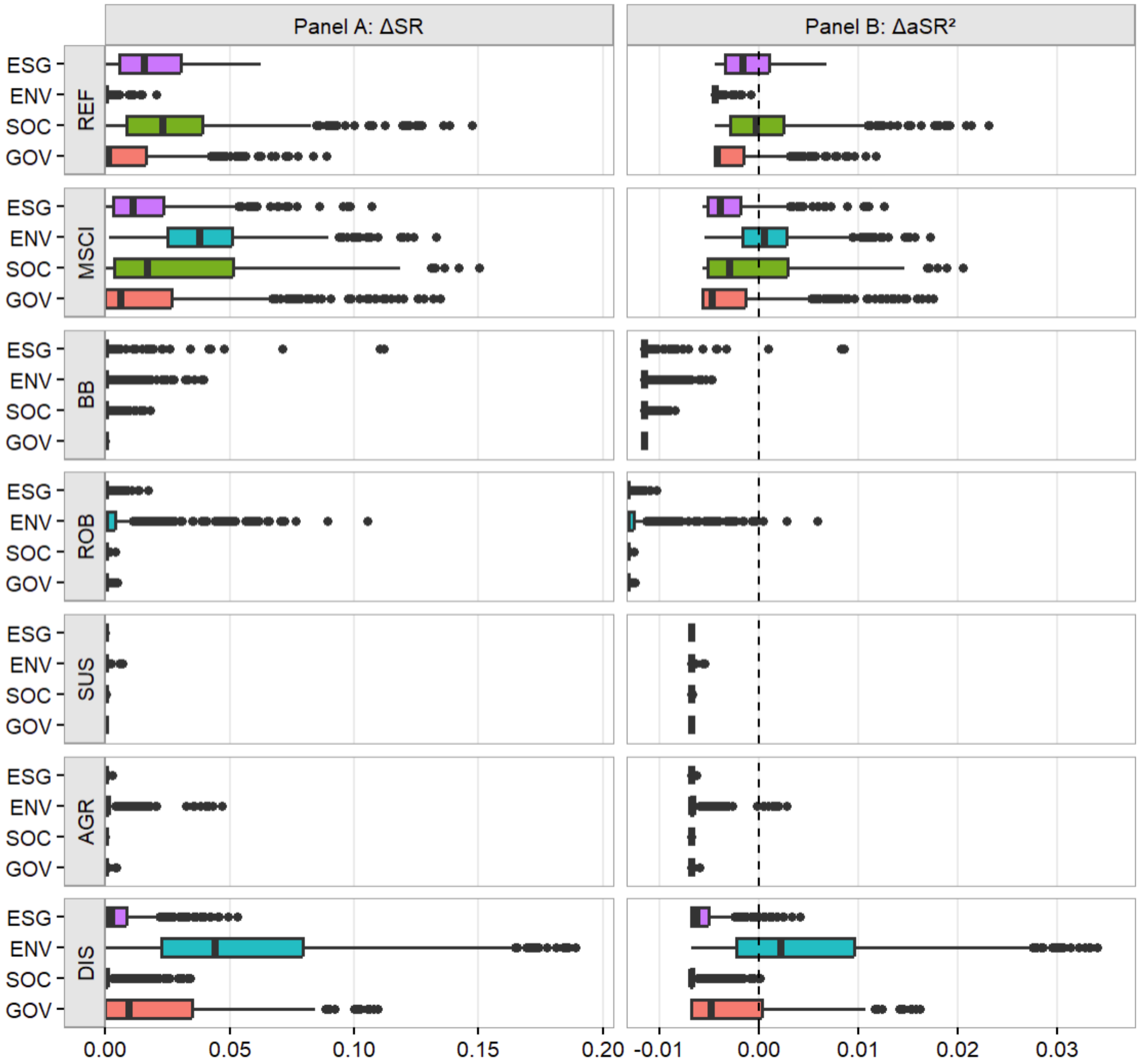
Panel B: Sharpe Ratios

Explanation: This figure depicts the boxplots of our rating factor returns (Panel A) and Sharpe ratios (Panel B) for ratings issued by one of the five major providers namely Refinitiv (REF; July 2003–June 2024), MSCI (July 2008–June 2024), Bloomberg (BB; July 2016–June 2024), RobecoSAM (ROB; July 2017–June 2024), Sustainalytics (SUS; July 2010–June 2024) and our self-calculated rating Agreement (AGR; July 2010–June 2024) and Disagreement (DIS; July 2010–June 2024) scores. For each rating provider/rating Agreement/Disagreement, we examine four rating types: ESG, Environmental (ENV), Social (SOC) and Governance (GOV).

We normalize all ratings before constructing rating factors. Each factor is computed as the average return difference between the two High-Rating (Big-High and Small-High) portfolios and the two Low-Rating (Big-Low and Small-Low) portfolios, as outlined in section 4.3.1. We create 512 factor specifications for each provider and each rating type. The boxplots show for each factor the median, 25th and 75th percentiles, as well as the minimum and maximum of the specifications.

Interpretation: Refinitiv's four factors and MSCI's three factors (except ENV) generate positive returns and positive Sharpe ratios. Bloomberg, RobecoSAM, Sustainalytics, and our AGR/DIS factor performance is generally neutral or negative, suggesting that rating-based factors (except REF) do not deliver consistent positive returns/risk-adjusted returns, and factor performance varies across providers.

Figure 6. Ex-post max Sharpe ratios



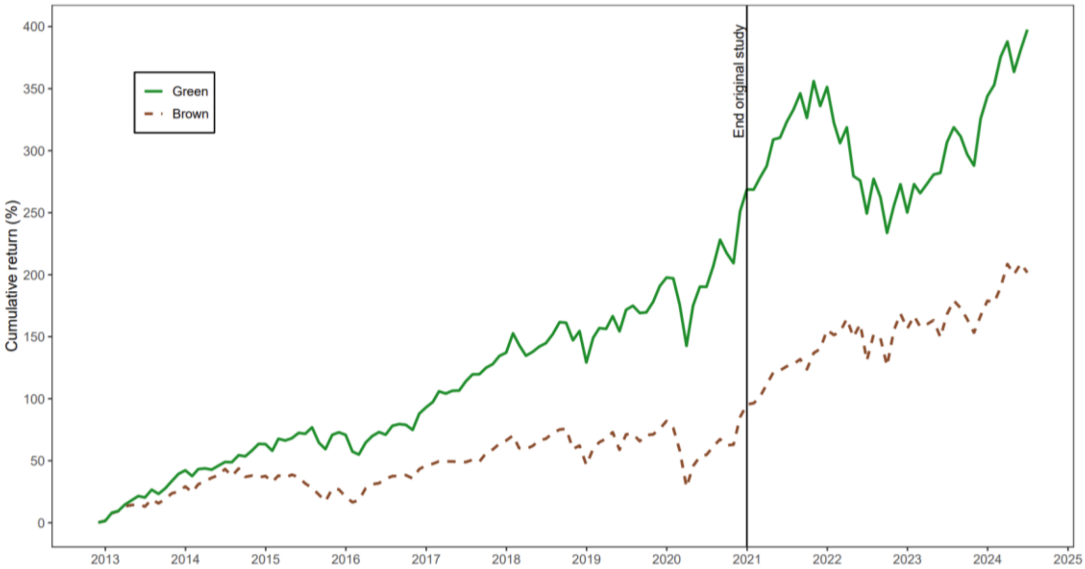
Explanation: This figure shows the results of the ex-post max Sharpe ratio analysis. Panel A (left) shows the change in the Sharpe ratio (ΔSR) when a sustainability rating factor is added to the BASE factor set. Panel B (right) shows the change in the bias-adjusted squared Sharpe ratio (ΔaSR²). As suggested by Hanauer (2020), the ΔaSR² corrects each model's squared Sharpe ratio for small-sample bias under joint normality. For this purpose, we first multiply our squared Sharpe ratio by *(T − K − 2)/T* and subtract *K/T* afterwards. Here, *K* is the number of factors while *T* is the number of return observations.

The BASE model is the Fama-French six-factor (FF6) benchmark. The added rating is either ESG, ENV, SOC, or GOV assigned by one or the five providers namely Refinitiv (REF; July 2003–June 2024), MSCI (July 2008–June 2024), Bloomberg (BB; July 2016–June 2024), RobecoSAM (ROB; July 2017–June 2024), Sustainalytics (SUS; July 2010–June 2024), or our self-created rating Agreement (AGR; July 2010–June 2024) and Disagreement (DIS; July 2010–June 2024). The boxplots summarize the distribution across 512 factor specifications. Within a provider, values are computed specification-by-specification relative to that provider's BASE model, so the dashed vertical line at zero is the no-change reference. Boxes to the right (left) of the dashed line indicate improvements (deterioration) versus the BASE model.

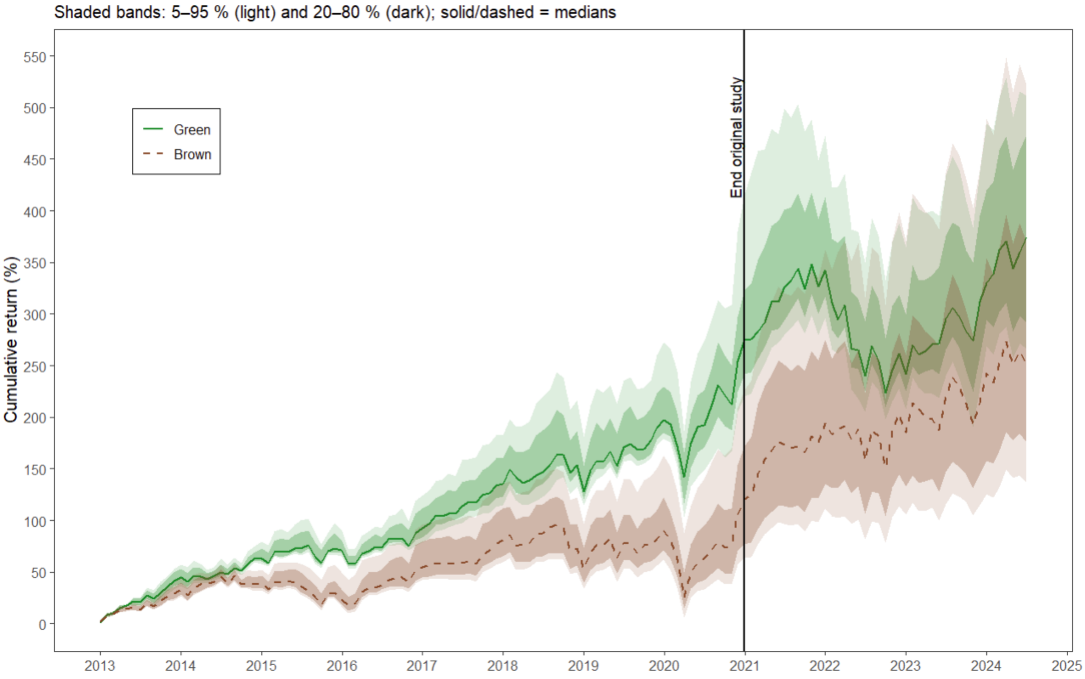
Interpretation: Adding a rating factor to the FF6 BASE set yields at best very small ex-post Sharpe ratio improvement (ΔSR), with only a few upper-tail or isolated cases (e.g. DIS's ENV, REF's SOC) reaching "higher" values. The changes in bias-adjusted squared-Sharpe ratios (ΔaSR²) are dominant by negative values, with few modest positive medians (e.g. DIS's ENV), and no consistent, statistically or economically meaningful improvement versus the FF6 BASE model.

Appendix A. Pástor et al. (2022) Replication: Green and Brown portfolios returns derived from their single-specification method and our multi-specification method

Panel A: Value-weighted green and brown portfolio returns, November 2012-June 2024



Panel B: Green and Brown portfolio returns across multi-specifications, November 2012-June 2024

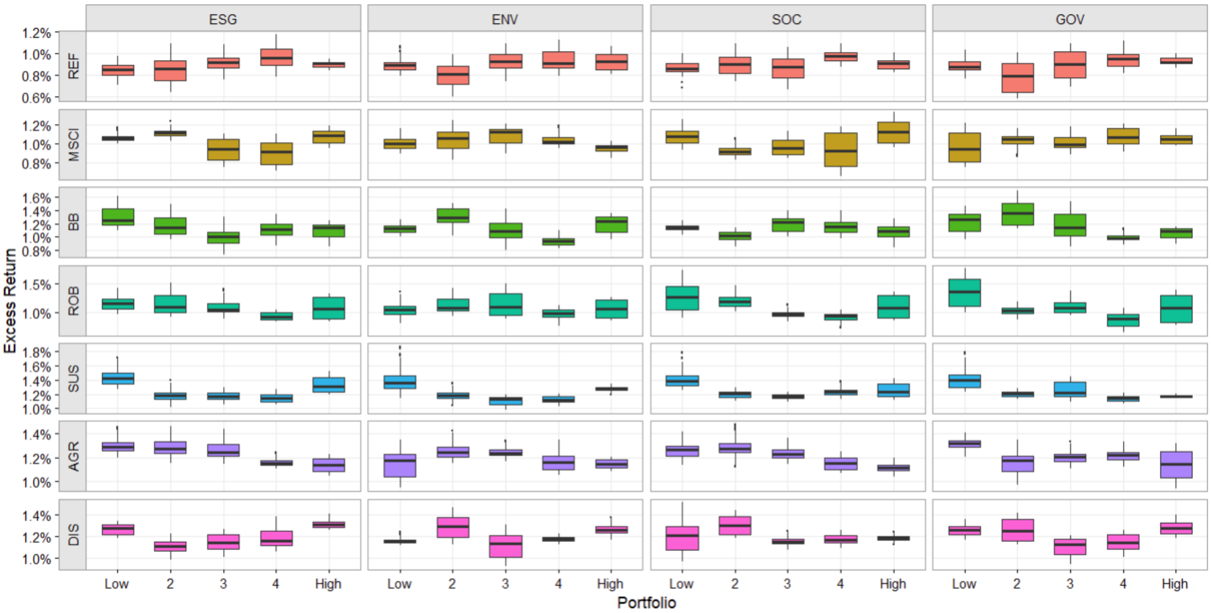


Explanation: Panel A of this Appendix features the value-weighted returns on green and brown portfolios created using the single-specification method of Pástor et al. (2022) over the extended period November 2012-June 2024. The cumulative return difference (CRD) between the two portfolios was 210%. Panel B shows the ranges of returns on green and brown portfolios constructed using our multi-specification method over the extended period November 2012-June 2024. We estimate 512 specifications of green and 512 specifications of brown portfolios. Over the original study period ending in December 2020, our CRDs varied between 61.5% and 321% with the median (mean) CRD of 164% (163%). Over the extended period ending in June 2024, our CRDs ranged between -78.4% and 268% with the median (mean) of 136% (110%). In the 5%-95% range during the extended period, we obtain CRDs of between -36% and 236%.

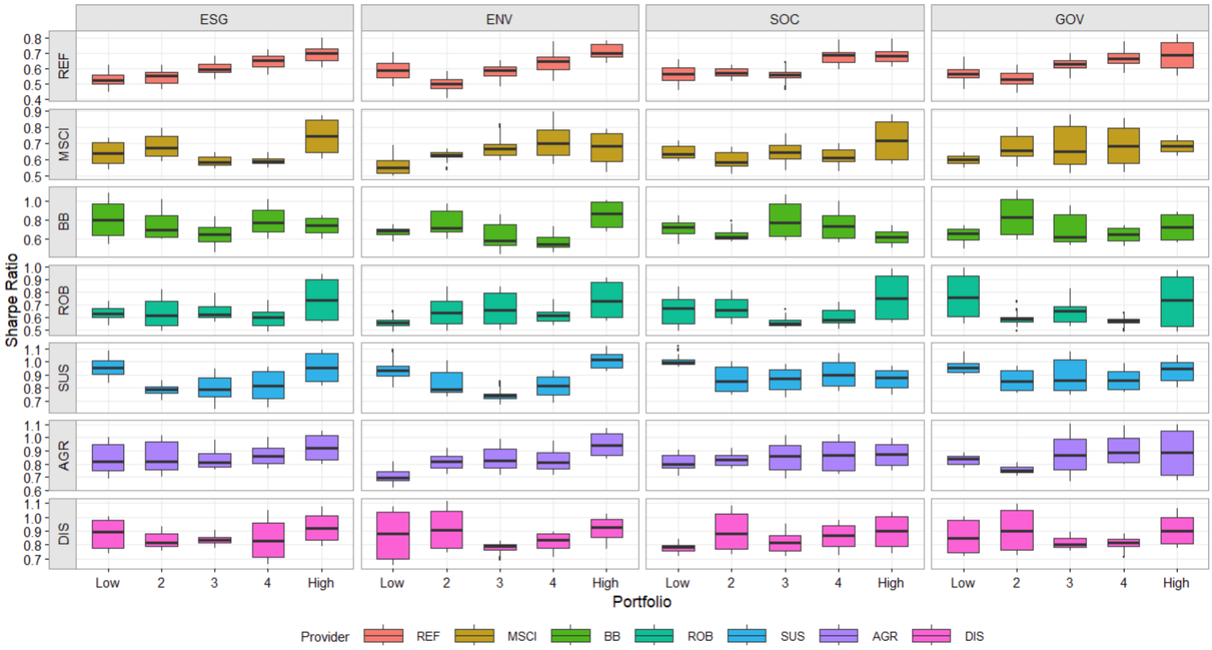
Interpretation: The single-specification replication shows clear green portfolio outperformance; however, our multi-specification estimates indicate that return difference between green and brown portfolios is sensitive to the chosen specification, especially after the original study period ends. Post-December 2020, we do not observe clear pattern of green portfolio outperformance across multi-specifications.

Appendix B. Single-sorted quintile portfolios' excess returns and Sharpe ratios

Panel A: Excess returns (%)



Panel B: Sharpe ratios

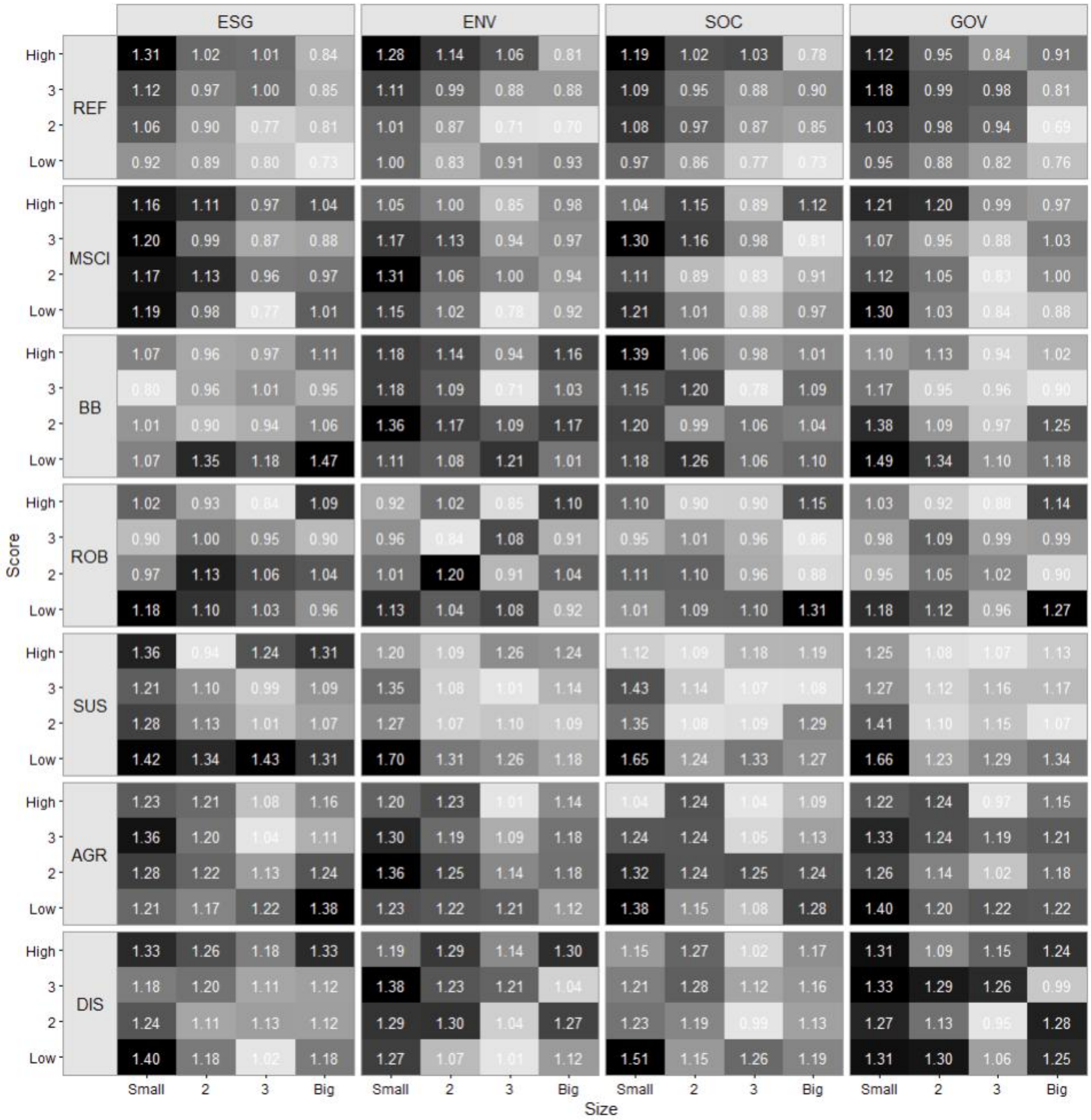


Explanation: This Appendix depicts the boxplots of excess returns (Panel A) and Sharpe ratios (Panel B) for single-sorted quintile 5x1 portfolios. For each rating provider, we use the rating scores in December of year $t_{(-1)}$ to construct the portfolio in June of year t_0 for the period from July of year t_0 to June of year t_1. We conduct analysis for each of five rating providers namely Refinitiv (REF; July 2003–June 2024), MSCI (July 2008–June 2024), Bloomberg (BB; July 2016–June 2024), RobecoSAM (ROB; July 2017–June 2024), Sustainalytics (SUS; July 2010–June 2024), and our self-constructed rating Agreement (AGR; July 2010–June 2024) and Disagreement (DIS; July 2010–June 2024) scores. For each rating provider/rating Agreement/Disagreement, we examine four rating categories: ESG, Environmental (ENV), Social (SOC) and Governance (GOV). We create 640 quantile portfolios (5 portfolios times 128 specifications each) for each provider and each rating type. The boxplots show for each portfolio the median, 25%- and 75%-quartile as well as the minimum and maximum of the specifications.

Interpretation: Overall, there is no consistent positive relationship between sustainability ratings and performance. Median excess returns and Sharpe ratios do not monotonically increase from low- to high-rated quintiles but instead fluctuate (and sometimes decline) across providers and rating types, suggesting that higher ratings are not consistently associated with better (risk-adjusted) returns.

Appendix C. Double-sorted 4x4 portfolios: excess returns and Sharpe ratios

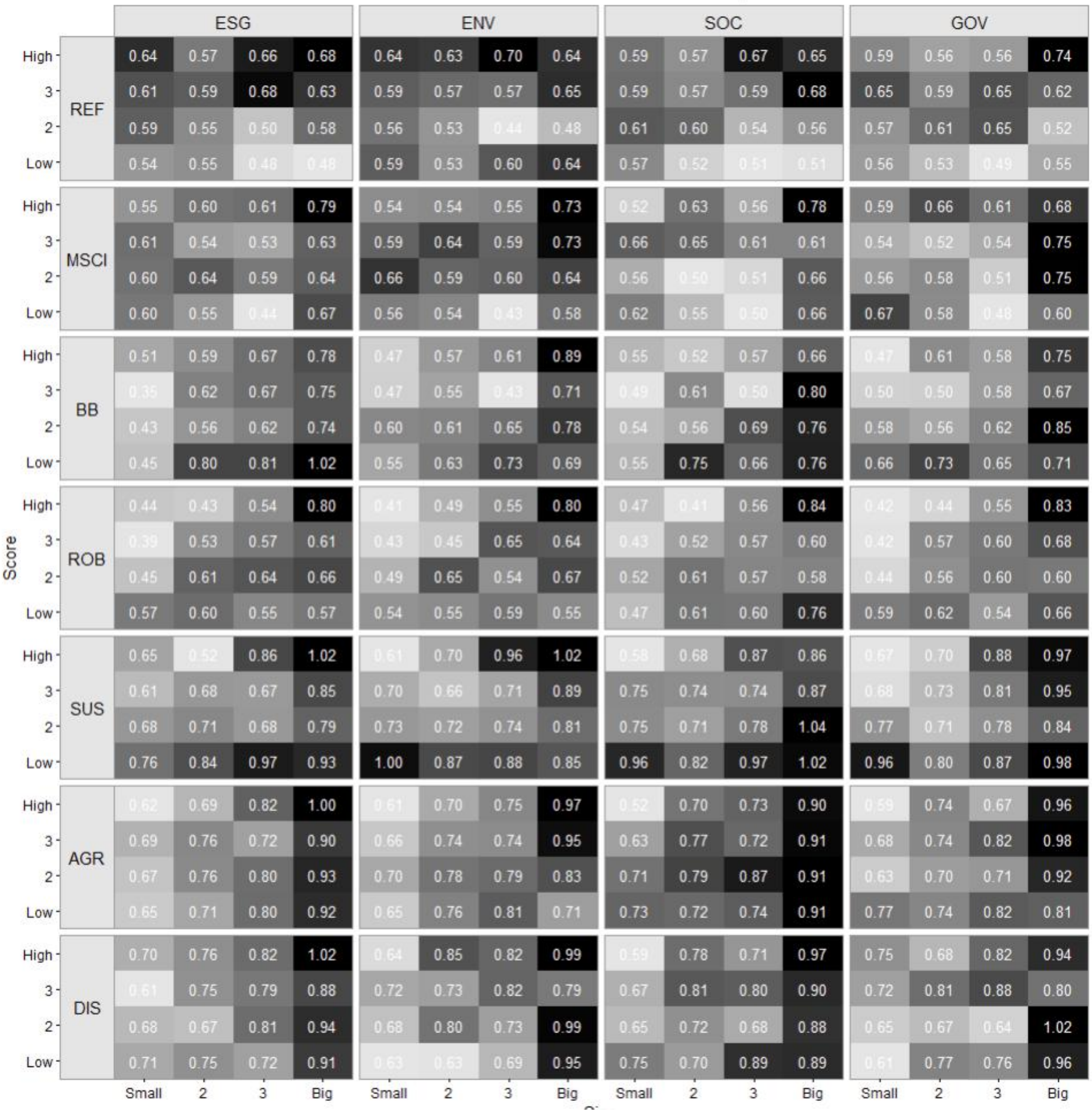
Panel A: Excess Returns (%)



		ESG				ENV				SOC				GOV			
	Score	Small	2	3	Big	Small	2	3	Big	Small	2	3	Big	Small	2	3	Big
REF	High	1.31	1.02	1.01	0.84	1.28	1.14	1.06	0.81	1.19	1.02	1.03	0.78	1.12	0.95	0.84	0.91
	3	1.12	0.97	1.00	0.85	1.11	0.99	0.88	0.88	1.09	0.95	0.88	0.90	1.18	0.99	0.98	0.81
	2	1.06	0.90	0.77	0.81	1.01	0.87	0.71	0.70	1.08	0.97	0.87	0.85	1.03	0.98	0.94	0.69
	Low	0.92	0.89	0.80	0.73	1.00	0.83	0.91	0.93	0.97	0.86	0.77	0.73	0.95	0.88	0.82	0.76
MSCI	High	1.16	1.11	0.97	1.04	1.05	1.00	0.85	0.98	1.04	1.15	0.89	1.12	1.21	1.20	0.99	0.97
	3	1.20	0.99	0.87	0.88	1.17	1.13	0.94	0.97	1.30	1.16	0.98	0.81	1.07	0.95	0.88	1.03
	2	1.17	1.13	0.96	0.97	1.31	1.06	1.00	0.94	1.11	0.89	0.83	0.91	1.12	1.05	0.83	1.00
	Low	1.19	0.98	0.77	1.01	1.15	1.02	0.78	0.92	1.21	1.01	0.88	0.97	1.30	1.03	0.84	0.88
BB	High	1.07	0.96	0.97	1.11	1.18	1.14	0.94	1.16	1.39	1.06	0.98	1.01	1.10	1.13	0.94	1.02
	3	0.80	0.96	1.01	0.95	1.18	1.09	0.71	1.03	1.15	1.20	0.78	1.09	1.17	0.95	0.96	0.90
	2	1.01	0.90	0.94	1.06	1.36	1.17	1.09	1.17	1.20	0.99	1.06	1.04	1.38	1.09	0.97	1.25
	Low	1.07	1.35	1.18	1.47	1.11	1.08	1.21	1.01	1.18	1.26	1.06	1.10	1.49	1.34	1.10	1.18
ROB	High	1.02	0.93	0.84	1.09	0.92	1.02	0.85	1.10	1.10	0.90	0.90	1.15	1.03	0.92	0.88	1.14
	3	0.90	1.00	0.95	0.90	0.96	0.84	1.08	0.91	0.95	1.01	0.96	0.86	0.98	1.09	0.99	0.99
	2	0.97	1.13	1.06	1.04	1.01	1.20	0.91	1.04	1.11	1.10	0.96	0.88	0.95	1.05	1.02	0.90
	Low	1.18	1.10	1.03	0.96	1.13	1.04	1.08	0.92	1.01	1.09	1.10	1.31	1.18	1.12	0.96	1.27
SUS	High	1.36	0.94	1.24	1.31	1.20	1.09	1.26	1.24	1.12	1.09	1.18	1.19	1.25	1.08	1.07	1.13
	3	1.21	1.10	0.99	1.09	1.35	1.08	1.01	1.14	1.43	1.14	1.07	1.08	1.27	1.12	1.16	1.17
	2	1.28	1.13	1.01	1.07	1.27	1.07	1.10	1.09	1.35	1.08	1.09	1.29	1.41	1.10	1.15	1.07
	Low	1.42	1.34	1.43	1.31	1.70	1.31	1.26	1.18	1.65	1.24	1.33	1.27	1.66	1.23	1.29	1.34
AGR	High	1.23	1.21	1.08	1.16	1.20	1.23	1.01	1.14	1.04	1.24	1.04	1.09	1.22	1.24	0.97	1.15
	3	1.36	1.20	1.04	1.11	1.30	1.19	1.09	1.18	1.24	1.24	1.05	1.13	1.33	1.24	1.19	1.21
	2	1.28	1.22	1.13	1.24	1.36	1.25	1.14	1.18	1.32	1.24	1.25	1.24	1.26	1.14	1.02	1.18
	Low	1.21	1.17	1.22	1.38	1.23	1.22	1.21	1.12	1.38	1.15	1.08	1.28	1.40	1.20	1.22	1.22
DIS	High	1.33	1.26	1.18	1.33	1.19	1.29	1.14	1.30	1.15	1.27	1.02	1.17	1.31	1.09	1.15	1.24
	3	1.18	1.20	1.11	1.12	1.38	1.23	1.21	1.04	1.21	1.28	1.12	1.16	1.33	1.29	1.26	0.99
	2	1.24	1.11	1.13	1.12	1.29	1.30	1.04	1.27	1.23	1.19	0.99	1.13	1.27	1.13	0.95	1.28
	Low	1.40	1.18	1.02	1.18	1.27	1.07	1.01	1.12	1.51	1.15	1.26	1.19	1.31	1.30	1.06	1.25

Appendix C. Double-sorted 4x4 portfolios: excess returns and Sharpe ratios (continued)

Panel B: Sharpe ratios



		ESG				ENV				SOC				GOV			
		Small	2	3	Big	Small	2	3	Big	Small	2	3	Big	Small	2	3	Big
REF	High	0.64	0.57	0.66	0.68	0.64	0.63	0.70	0.64	0.59	0.57	0.67	0.65	0.59	0.56	0.56	0.74
	3	0.61	0.59	0.68	0.63	0.59	0.57	0.57	0.65	0.59	0.57	0.59	0.68	0.65	0.59	0.65	0.62
	2	0.59	0.55	0.50	0.58	0.56	0.53	0.44	0.48	0.61	0.60	0.54	0.56	0.57	0.61	0.65	0.52
	Low	0.54	0.55	0.46	0.48	0.59	0.53	0.60	0.64	0.57	0.52	0.51	0.51	0.56	0.53	0.49	0.55
MSCI	High	0.55	0.60	0.61	0.79	0.54	0.54	0.55	0.73	0.52	0.63	0.56	0.78	0.59	0.66	0.61	0.68
	3	0.61	0.54	0.53	0.63	0.59	0.64	0.59	0.73	0.66	0.65	0.61	0.61	0.54	0.52	0.54	0.75
	2	0.60	0.64	0.59	0.64	0.66	0.59	0.60	0.64	0.56	0.50	0.51	0.66	0.56	0.58	0.51	0.75
	Low	0.60	0.55	0.44	0.67	0.56	0.54	0.43	0.58	0.62	0.55	0.50	0.66	0.67	0.58	0.48	0.60
BB	High	0.51	0.59	0.67	0.78	0.47	0.57	0.61	0.89	0.55	0.52	0.57	0.66	0.47	0.61	0.58	0.75
	3	0.35	0.62	0.67	0.75	0.47	0.55	0.43	0.71	0.49	0.61	0.50	0.80	0.50	0.50	0.58	0.67
	2	0.43	0.56	0.62	0.74	0.60	0.61	0.65	0.78	0.54	0.56	0.69	0.76	0.58	0.56	0.62	0.85
	Low	0.45	0.80	0.81	1.02	0.55	0.63	0.73	0.69	0.55	0.75	0.66	0.76	0.66	0.73	0.65	0.71
ROB	High	0.44	0.43	0.54	0.80	0.41	0.49	0.55	0.80	0.47	0.41	0.56	0.84	0.42	0.44	0.55	0.83
	3	0.39	0.53	0.57	0.61	0.43	0.45	0.65	0.64	0.43	0.52	0.57	0.60	0.42	0.57	0.60	0.68
	2	0.45	0.61	0.64	0.66	0.49	0.65	0.54	0.67	0.52	0.61	0.57	0.58	0.44	0.56	0.60	0.60
	Low	0.57	0.60	0.55	0.57	0.54	0.55	0.59	0.55	0.47	0.61	0.60	0.76	0.59	0.62	0.54	0.66
SUS	High	0.65	0.52	0.86	1.02	0.61	0.70	0.96	1.02	0.58	0.68	0.87	0.86	0.67	0.70	0.88	0.97
	3	0.61	0.68	0.67	0.85	0.70	0.66	0.71	0.89	0.75	0.74	0.74	0.87	0.68	0.73	0.81	0.95
	2	0.68	0.71	0.68	0.79	0.73	0.72	0.74	0.81	0.75	0.71	0.78	1.04	0.77	0.71	0.78	0.84
	Low	0.76	0.84	0.97	0.93	1.00	0.87	0.88	0.85	0.96	0.82	0.97	1.02	0.96	0.80	0.87	0.98
AGR	High	0.62	0.69	0.82	1.00	0.61	0.70	0.75	0.97	0.52	0.70	0.73	0.90	0.59	0.74	0.67	0.96
	3	0.69	0.76	0.72	0.90	0.66	0.74	0.74	0.95	0.63	0.77	0.72	0.91	0.68	0.74	0.82	0.98
	2	0.67	0.76	0.80	0.93	0.70	0.78	0.79	0.83	0.71	0.79	0.87	0.91	0.63	0.70	0.71	0.92
	Low	0.65	0.71	0.80	0.92	0.65	0.76	0.81	0.71	0.73	0.72	0.74	0.91	0.77	0.74	0.82	0.81
DIS	High	0.70	0.76	0.82	1.02	0.64	0.85	0.82	0.99	0.59	0.78	0.71	0.97	0.75	0.68	0.82	0.94
	3	0.61	0.75	0.79	0.88	0.72	0.73	0.82	0.79	0.67	0.81	0.80	0.90	0.72	0.81	0.88	0.80
	2	0.68	0.67	0.81	0.94	0.68	0.80	0.73	0.99	0.65	0.72	0.68	0.88	0.65	0.67	0.64	1.02
	Low	0.71	0.75	0.72	0.91	0.63	0.63	0.69	0.95	0.75	0.70	0.89	0.89	0.61	0.77	0.76	0.96

Explanation: This Appendix depicts the median excess returns (Panel A) and Sharpe ratios (Panel B) for double-sorted 4x4 portfolios (sorted by market cap and rating scores). For each rating provider, we use the rating scores in December of year $t_{(-1)}$ to construct the portfolio in June of year t_0 for the period from July of year t_0 to June of year t_1. We conduct analysis for each of our five rating providers namely Refinitiv (July 2003–June 2024), MSCI (July 2008–June 2024), Bloomberg (BB; July 2016–June 2024), RobecoSAM (ROB; July 2017–June 2024), Sustainalytics (SUS; July 2010–June 2024), and our self-constructed rating Agreement (AGR; July 2010–June 2024) and Disagreement (DIS; July 2010–June 2024) score. For each rating provider/rating Agreement/Disagreement, we examine four rating types namely ESG, Environmental (ENV), Social (SOC) and Governance (GOV). A darker shade in the heatmap indicates a higher median excess return (Panel A) or a greater Sharpe ratio (Panel B).

Interpretation: Median excess returns and Sharpe ratios of double-sorted 4×4 rating×size portfolios show no consistent monotonic increase as rating score improves. The relationships between rating scores and excess returns/Sharpe ratios differ across providers and rating types; though to some extent, they are affected by size.

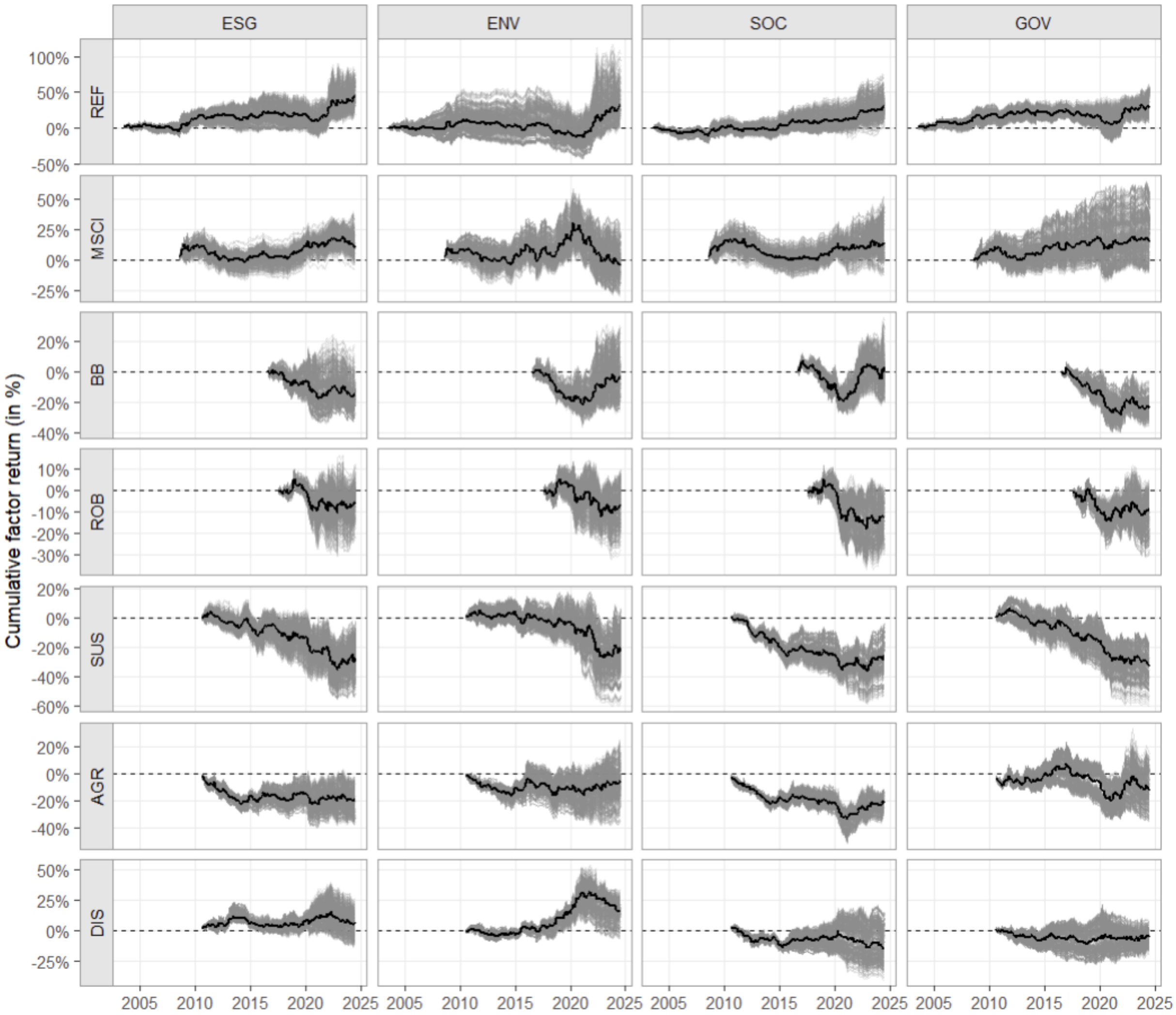
Appendix D. Differences in median factor returns due to alternative decisions

	Weighting	**Market cap**	**Price**	**Excl. neg. Book Value**	**Excl. neg. Earnings**	**Include Finance**	**Include Utilities**	**Exchange**	**Sorting decision**
Rating	Value – equal-weighted	Min $300m – $0	Min $5 – $0	Yes – No	Yes – No	Yes – No	Yes – No	ALL – NYSE	Dep. – Indep.
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: Refinitiv (July 2003 – June 2024)									
ESG	-0.04	0.00	-0.02	0.01	0.01	0.02	0.00	0.01	-0.06
ENV	-0.06	-0.01	-0.01	0.02	-0.01	-0.06	-0.03	0.03	-0.01
SOC	-0.05	0.01	0.00	0	-0.02	0.01	0.01	0.01	0.01
GOV	0.03	-0.01	-0.01	0.02	0.03	0.02	0.03	0.02	-0.02
Panel B: MSCI (July 2008 – June 2024)									
ESG	0.03	0.01	-0.02	-0.01	0.02	0.01	0.03	-0.03	-0.02
ENV	-0.05	0.00	0.00	0.04	0.05	0.11	0.00	-0.01	-0.01
SOC	0.05	0.00	0.00	-0.01	0.03	0.01	-0.03	0.00	0.00
GOV	0.12	0.00	-0.01	0.01	0.05	-0.08	0.03	0.01	-0.03
Panel C: Bloomberg (July 2016 – June 2024)									
ESG	-0.15	0.00	0.00	0.00	-0.16	-0.07	-0.10	-0.01	0.03
ENV	0.26	-0.01	0.00	0.01	-0.14	-0.02	-0.01	-0.06	-0.04
SOC	0.13	-0.01	-0.01	0.00	-0.05	0.04	-0.11	0.01	0.07
GOV	0.04	0.01	-0.02	0.03	0.21	0.11	-0.02	0.05	-0.12
Panel D: RobecoSAM (July 2017 – June 2024)									
ESG	0.15	-0.01	0.00	-0.03	-0.05	0.04	0.06	0.02	0.03
ENV	0.17	0.02	0.00	-0.04	-0.07	0.07	0.01	0.03	0.06
SOC	0.04	0.02	-0.01	-0.04	0.09	-0.03	0.05	0.00	0.02
GOV	-0.03	-0.01	0.00	-0.02	0.01	0.03	0.03	-0.03	0.15
Panel E: Sustainalytics (July 2010 – June 2024)									
ESG	0.01	0.04	0.01	0.00	0.03	0.05	0.03	-0.01	0.04
ENV	-0.04	0.00	0.00	0.01	0.06	0.08	0.03	-0.01	0.02
SOC	0.00	0.01	0.02	0.00	0.09	-0.03	0.04	0.00	-0.03
GOV	0.11	0.00	0.01	0.02	0.09	0.06	-0.05	0.00	-0.03
Panel F: Agreement (July 2010 – June 2024)									
ESG	-0.02	0.00	0.00	-0.02	0.00	-0.01	0.00	0.01	0.08
ENV	0.00	0.01	0.00	-0.04	-0.03	0.06	0.01	-0.01	0.08
SOC	0.01	0.01	0.00	0.00	0.06	-0.03	0.03	0.02	0.04
GOV	0.09	0.00	0.00	0.03	-0.02	0.06	-0.04	-0.01	0.03
Panel G: Disagreement (July 2010 – June 2024)									
ESG	-0.01	0.00	0.01	0.03	-0.01	0.03	-0.03	0.02	0.07
ENV	0.06	0.00	0.00	0.04	0.05	0.03	0.00	0.00	-0.01
SOC	0.14	0.00	0.00	0.02	0.08	0.05	0.01	0.02	-0.01
GOV	-0.11	-0.01	-0.01	0.00	0.00	-0.04	-0.01	0.01	0.00

Explanation: This Appendix shows the differences in median factor returns that can be attributed to the decisions made at one node at a time, holding specifications relating to other nodes in Figure 1 constant. We create 512 factor specifications for each provider and each rating type. Each factor is computed as the average return difference between the two High-Rating (Big-High and Small-High) portfolios and the two Low-Rating (Big-Low and Small-Low) portfolios, as outlined in section 4.3.1. We report the median factor return difference due to weighting schemes (column 2), market cap thresholds (column 3), price thresholds (column 4), excluding (Yes) vs including (No) negative book value (column 5) and negative earnings per share (column 6), sector inclusion (Yes) vs exclusion (No) (columns 7-8), stock exchanges (column 9), and sorting procedure (column 10). For example, the return difference between 256 value-weighted ESG factor specifications and 256 equally weighted ESG factor specifications (column 2 and row 1 in each Panel) is only due to the weighting decision as these two sets of 256 factor specifications are identical except the weighting scheme.

Interpretation: Decisions concerning weighting scheme (column 2), excluding vs. including firms with negative earnings (column 6), and including vs. excluding Finance sector (column 7) result in noticeable median factor return differences for some providers.

Appendix E. Rating factor performance over time



Explanation: This Appendix depicts the rating factor performance over time for the five rating providers, Refinitiv (REF; July 2003–June 2024), MSCI (July 2008–June 2024), Bloomberg (BB; July 2016–June 2024), RobecoSAM (ROB; July 2017–June 2024), Sustainalytics (SUS; July 2010–June 2024), and our self-calculated rating Agreement (AGR; July 2010–June 2024) and rating Disagreement (DIS; July 2010–June 2024) scores. For each rating provider/rating Agreement/Disagreement, we examine four rating types: ESG, Environmental (ENV), Social (SOC) and Governance (GOV). We create 512 factor specifications for each provider and each rating. Each factor is computed as the average return difference between the two High-Rating (Big-High and Small-High) portfolios and the two Low-Rating (Big-Low and Small-Low) portfolios, as outlined in section 4.3.1. In each graph, the thick black line shows the median return across 512 factor specifications, while a grey line shows the return of one factor specification.

Interpretation: Median factor returns show no clear upward trend as time passes. We observe mostly fluctuating or declining returns, except for REF's SOC factor. Several factors (notably SUS, AGR, DIS, and some ROB/BB factors) display persistent (large) negative returns during most or throughout the study period.

Appendix F. Summary statistics of cumulative factor returns across specifications as of June 2024

Rating	Mean	Median	Min	Max	Max-Min	Q5	Q10	Q20	Q80	Q90	Q95
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Panel A: Refinitiv											
ESG	45.0%	44.7%	18.8%	79.0%	60.2%	25.3%	28.7%	33.4%	56.8%	63.0%	66.5%
ENV	33.3%	31.0%	1.7%	102.0%	100.3%	7.4%	13.5%	18.5%	46.7%	57.0%	65.1%
SOC	31.9%	30.0%	3.2%	72.9%	69.7%	14.1%	16.7%	21.0%	42.7%	49.0%	55.5%
GOV	32.3%	29.1%	7.1%	61.3%	54.2%	13.2%	15.9%	19.5%	47.5%	51.5%	54.1%
Panel B: MSCI											
ESG	11.6%	10.5%	-7.2%	35.4%	42.6%	0.6%	2.7%	5.4%	18.1%	21.7%	24.5%
ENV	-2.0%	-2.8%	-28.5%	28.8%	57.3%	-22.4%	-17.9%	-13.3%	10.0%	14.5%	19.5%
SOC	16.2%	12.7%	-13.5%	52.8%	66.3%	-2.3%	0.4%	4.5%	29.8%	36.5%	42.1%
GOV	16.6%	14.7%	-11.4%	54.9%	66.3%	-7.2%	-5.5%	-0.4%	32.2%	41.2%	48.6%
Panel C: Bloomberg											
ESG	-14.8%	-15.0%	-30.5%	16.9%	47.4%	-26.9%	-24.3%	-22.7%	-9.4%	-3.5%	0.5%
ENV	-2.9%	-4.1%	-26.0%	30.9%	56.9%	-21.5%	-19.5%	-16.0%	8.8%	15.5%	20.5%
SOC	0.7%	0.1%	-21.0%	33.3%	54.3%	-17.0%	-13.7%	-9.0%	10.2%	16.4%	20.7%
GOV	-23.3%	-23.8%	-36.0%	-6.9%	29.1%	-31.6%	-30.0%	-28.2%	-18.5%	-16.3%	-13.7%
Panel D: RobecoSAM											
ESG	-5.5%	-4.9%	-23.0%	12.3%	35.3%	-16.4%	-13.9%	-11.4%	0.1%	2.4%	4.6%
ENV	-6.8%	-6.2%	-29.0%	9.7%	38.7%	-20.3%	-17.7%	-14.9%	0.6%	4.7%	6.3%
SOC	-12.6%	-12.4%	-30.5%	2.1%	32.6%	-23.8%	-21.7%	-19.1%	-6.3%	-3.5%	-1.6%
GOV	-10.0%	-8.7%	-29.3%	5.3%	34.6%	-22.8%	-20.6%	-16.5%	-3.6%	-1.3%	0.3%
Panel E: Sustainalytics											
ESG	-26.9%	-27.4%	-53.0%	-2.2%	50.8%	-45.4%	-42.1%	-35.6%	-16.4%	-12.9%	-10.7%
ENV	-21.0%	-20.2%	-58.0%	6.3%	64.3%	-45.9%	-37.2%	-31.8%	-10.1%	-3.8%	-1.2%
SOC	-27.1%	-26.5%	-53.6%	-3.7%	49.9%	-46.7%	-38.2%	-35.3%	-20.3%	-16.3%	-9.9%
GOV	-32.8%	-31.9%	-60.2%	-16.2%	44.0%	-48.5%	-45.1%	-40.1%	-25.0%	-20.2%	-18.0%

Appendix F. Summary statistics of cumulative factor returns across specifications as of June 2024 (continued)

Rating (1)	Mean (2)	Median (3)	Min (4)	Max (5)	Max-Min (6)	Q5 (7)	Q10 (8)	Q20 (9)	Q80 (10)	Q90 (11)	Q95 (12)
Panel F: Agreement											
ESG	-19.2%	-18.6%	-37.2%	-2.4%	34.8%	-33.1%	-31.7%	-28.3%	-10.6%	-7.8%	-5.8%
ENV	-4.9%	-4.7%	-35.2%	26.4%	61.6%	-25.8%	-19.5%	-14.9%	4.8%	11.8%	16.3%
SOC	-21.5%	-21.5%	-32.6%	-12.3%	20.3%	-28.7%	-27.3%	-25.5%	-17.5%	-15.4%	-14.4%
GOV	-10.3%	-12.1%	-34.7%	14.9%	49.6%	-30.5%	-27.2%	-21.1%	2.5%	4.3%	6.9%
Panel G: Disagreement											
ESG	5.8%	6.0%	-11.8%	26.2%	38.0%	-7.0%	-3.8%	-0.3%	11.4%	15.3%	18.6%
ENV	15.6%	15.3%	-5.8%	33.7%	39.5%	0.6%	3.5%	10.0%	22.8%	26.2%	28.0%
SOC	-12.2%	-15.0%	-40.3%	16.1%	56.4%	-30.5%	-28.0%	-24.7%	1.6%	5.6%	9.0%
GOV	-4.8%	-3.8%	-19.6%	8.7%	28.3%	-15.1%	-13.1%	-9.7%	-0.6%	1.4%	2.9%

Explanation: This table reports the statistics of end-of-sample (June 2024) cumulative factor returns (see Appendix E) across specifications for our five providers Refinitiv (REF; July 2003–June 2024), MSCI (July 2008–June 2024), Bloomberg (BB; July 2016–June 2024), RobecoSAM (ROB; July 2017–June 2024), Sustainalytics (SUS; July 2010–June 2024) and our self-calculated rating Agreement (AGR; July 2010–June 2024) and Disagreement (DIS; July 2010–June 2024) scores. For each rating provider/rating Agreement/Disagreement, we examine four rating types: ESG, Environmental (ENV), Social (SOC) and Governance (GOV). Each factor is computed as the average return difference between the two High-Rating (Big-High and Small-High) portfolios and the two Low-Rating (Big-Low and Small-Low) portfolios, as outlined in section 4.3.1.

Interpretation: Rating factor cumulative returns shift considerably across factor specifications for the same rating provider and rating dimension.

Online Appendix A. "Sustainable investing with ESG rating uncertainty" by Avramov et al. (*Journal of Financial Economics*, 2022) – Replication details

Online Appendix A.1. Sample

Our replication of Avramov et al. (2022) follows their sample construction by using NYSE, AMEX and NASDAQ common stocks with the share codes of 10 or 11 obtained from CRSP. Avramov et al. (2022) utilize six ESG ratings: Refinitiv's Combined ESG score, MSCI IVA's ESG score, Bloomberg's ESG Disclosure Score, Sustainalytics Ranks, RobecoSAM Total Sustainability rank, and MSCI KLD. The rating sample of the original study starts in 2002 and ends in 2018. Except for MSCI KLD data which was last available in 2018, we use ESG ratings from the same five other providers in our replication. Considering the number of firms with available rating data, our rating sample starts in 2007 and ends in 2022.

Online Appendix Table A.1 compares the replication (R) and original study (O) samples. Panel A shows the number of stocks covered by each data vendor, and Panel B features the number of stocks covered by multiple data vendors. The difference between the replication and original samples is denoted as R-O. For both Panels A and B, we use December ESG ratings assigned by each provider.[48]

Online Appendix Table A.1. Number of stocks over time
(Original study Table B.1 Number of Stocks Covered By Each Data Vendor (Online Appendix: A - 10))

Panel A: Number of stocks covered by each data vendor

Year	Refinitiv (1)			MSCI IVA (2)			Bloomberg (3)			Sustainalytics (4)			RobecoSAM (5)		
	R	O	R-O	R	O	R-O	R	O	R-O	R	O	R-O	R	O	R-O
2002	404	398	6												
2003	431	400	31												
2004	565	535	30												
2005	606	600	6				113	125	-12						
2006	620	606	14		528	-528	217	209	8						
2007	656	620	36	597	609	-12	901	709	192						
2008	841	789	52	581	600	-19	1290	984	306						
2009	891	892	-1	566	599	-33	1369	1065	304	484		484			
2010	911	915	-4	532	551	-19	2334	1957	377	700		700			
2011	900	912	-12	513	537	-24	2520	2077	443	747		747			
2012	883	895	-12	2072	2253	-181	2622	2149	473	763		763			
2013	887	890	-3	2233	2388	-155	2663	2242	421	623		623			
2014	966	885	81	2267	2328	-61	2732	2380	352	733	413	320			
2015	1574	1436	138	2256	2282	-26	2767	2514	253	818	441	377			
2016	2154	2083	71	2278	2255	23	2715	2530	185	2074	460	1614	480	419	61
2017	2516	2218	298	2170	2139	31	2698	2658	40	2245	452	1793	630	616	14

[48] Using all ESG ratings available in a given year would have increased the sample size, but at the cost of overstating the actual number of rating observations employed.

Online Appendix Table A.1 Number of stocks over time (continued)

(Original study Table B.1 Number of Stocks Covered By Each Data Vendor (Online Appendix: A - 10))

Year	**Refinitiv** (1)			**MSCI IVA** (2)			**Bloomberg** (3)			**Sustainalytics** (4)			**RobecoSAM** (5)		
	R	O	R-O	R	O	R-O	R	O	R-O	R	O	R-O	R	O	R-O
2018	2635	2178	457	2139	2104	35	2721	2794	-73	2264	473	1791	753	818	-65
2019	2708			2182			2704			2365			1066		
2020	2803			2179			2694			2458			1078		
2021	2842			2241			2892			2663			1299		
2022	2867			2405			2800			2616			1279		

Panel A shows a comparison of our stock coverage versus Avramov et al. (2022) sample coverage. Columns (1-5) show the numbers of stocks with ratings assigned by Refinitiv, MSCI IVA, Bloomberg, Sustainalytics and RobecoSAM, respectively. Overall, our samples starting in 2007 (denoted as R) tracks their respective samples during 2007-2018 (denoted as O) quite closely. Ours often exceeds theirs, particularly for Sustainalytics ratings.[49]

Panel B: Number of stocks covered by multiple data vendors

Year	**N=1** (1)			**N=2** (2)			**N=3** (3)			**N=4** (4)			**N=5** (5)			**N ≥ 2** (6)		
	R	O	R-O	R	O	R-O	R	O	R-O	R	O	R-O	R	O	R-O	R	O	R-O
2002	404	677	-273	0	388	-388	0	0	0	0	0	0	0	0	0	0	388	-388
2003	431	2409	-1978	0	398	-398	0	0	0	0	0	0	0	0	0	0	398	-398
2004	565	2324	-1759	0	531	-531	0	0	0	0	0	0	0	0	0	0	531	-531
2005	551	2199	-1648	84	518	-434	0	59	-59	0	0	0	0	0	0	84	577	-493
2006	479	2069	-1590	179	241	-62	0	349	-349	0	100	-100	0	0	0	179	690	-511
2007	501	1756	-1255	216	380	-164	407	264	143	0	299	-299	0	0	0	623	943	-320
2008	589	1579	-990	340	505	-165	481	320	161	0	351	-351	0	0	0	821	1176	-355
2009	561	1601	-1040	295	487	-192	205	373	-168	386	365	21	0	0	0	886	1225	-339
2010	1470	1240	230	173	1093	-920	327	385	-58	420	368	52	0	0	0	920	1846	-926
2011	1643	1136	507	157	1109	-952	353	392	-39	416	367	49	0	0	0	926	1868	-942
2012	650	631	19	1183	702	481	204	1060	-856	678	625	53	0	0	0	2065	2387	-322
2013	673	741	-68	1286	591	695	283	1038	-755	578	652	-74	0	0	0	2147	2281	-134
2014	709	781	-72	1218	586	632	311	1030	-719	655	289	366	0	381	-381	2184	2286	-102
2015	735	851	-116	729	341	388	754	811	-57	740	669	71	0	431	-431	2223	2252	-29
2016	444	797	-353	353	645	-292	444	1119	-675	1251	87	1164	443	391	52	2491	2242	249
2017	243	781	-538	383	512	-129	502	1140	-638	1226	162	1064	568	442	126	2679	2256	423
2018	211	817	-606	411	425	-14	508	1042	-534	1145	336	809	675	446	229	2739	2249	490
2019	211			373			472			953			968			2766		
2020	227			380			464			972			989			2805		
2021	259			379			499			957			1119			2954		
2022	226			410			492			985			1101			2988		

Panel B shows a comparison of stocks covered by multiple raters. For example, N=3 means that a stock is covered by three rating providers. Column 6 shows the number of firms with ratings assigned by at least two providers, which determines the number of stocks available for our portfolio construction. Due to missing MSCI KLD data and considering the number of firms with available rating data, our rating sample (denoted as R) starts in 2007 (instead of 2002 as in the original study). From 2016 onwards our coverage becomes broader, mainly through Sustainalytics, and our number of firms with ratings from at least two providers exceeds that of the original study.

[49] Our Sustainalytics rating data begins in 2009, whereas Avramov et al. (2022) report data starting in 2014. To clarify the construction of their Sustainalytics Rank variable and related coding procedures, we contacted the corresponding author but did not receive a response. The *Journal of Financial Economics* office advised us that no replication files (data and programming code) for this paper is available.

Online Appendix A.2. Method

As in Avramov et al. (2022), we calculate percentiles ranks for each ESG rating score which was normalized between 0 and 1. Then, for each rater pair, we calculate the standard deviation and take its average across all pairs to obtain the firm-level ESG rating uncertainty. Similarly, we calculate the average rank for each rater pair and then take the mean across all pairs. Additionally, as a robustness test for ESG rating and ESG rating uncertainty, we calculate the mean rank and standard deviation of all ratings instead of rater pairs. Similar to Avramov et al. (2022), we require at least two data vendors to calculate the ESG rating uncertainty.

Online Appendix Table A.2 compares pairwise ESG rating uncertainty, ESG rating correlations, and the distribution of stock-level ESG characteristics between our samples and Avramov et al.'s (2022) samples. Our ratings were issued from 2007 (instead of 2002 as the original study) to 2018. For both Panels A and B, the replicated values are nearly identical to the original ones, with only minor deviations in specific provider pairs (most notably those involving Sustainalytics). Nevertheless, we obtain qualitatively similar statistics.

Panel C shows that the distributional properties of ESG ratings and ESG rating uncertainties (for PAIR and ALL measures) in our replication are closely aligned with those of the original sample, with differences that are quantitatively minor and economically negligible.

We employ the same approach as Avramov et al. (2022) to form portfolios. At the end of each year, we sort firms into ESG rating-uncertainty quintiles and, within each of these uncertainty quintiles, sort firms into ESG-rating quintiles (dependent sort). We then form value-weighted portfolios to compute monthly returns. Our portfolio of main interest is *Low LMH-R,* which is defined as the *Low ESG Rating* minus *High ESG rating* within the *Low Uncertainty* quintile.[50]

[50] The *Low Uncertainty* quintile consist of the 20% stocks with the lowest uncertainty rating. Within this quintile, there are five portfolios ranging from *Low ESG Rating* to *High ESG Rating* (Low, 2, 3, 4, High).

Online Appendix Table A.2. Summary statistics

(Original study Table B.3: Summary Statistics (Online Appendix: A - 12))

		Panel A: Pairwise ESG rating uncertainty (Jan 2008 – Dec 2019)			**Panel B: Pairwise ESG rating correlations (Jan 2008 – Dec 2019)**		
Rater 1	Rater 2	**Replication**	**Original**	**R – O**	**Replication**	**Original**	**R – O**
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
REF	MSCI	0.186	0.185	0.001	0.335	0.326	0.009
REF	BB	0.144	0.134	0.010	0.589	0.639	-0.050
REF	SUS	0.144	0.144	0.000	0.600	0.595	0.005
REF	ROB	0.149	0.149	0.000	0.571	0.547	0.024
MSCI	BB	0.206	0.195	0.011	0.223	0.253	-0.030
MSCI	SUS	0.185	0.171	0.014	0.356	0.411	-0.055
MSCI	ROB	0.180	0.181	-0.001	0.376	0.353	0.023
BB	SUS	0.138	0.133	0.005	0.613	0.677	-0.064
BB	ROB	0.142	0.138	0.004	0.609	0.645	-0.036
SUS	ROB	0.142	0.119	0.023	0.614	0.707	-0.093

Panel C: Quantile distribution of stock-level ESG characteristics (Jan 2008 – Dec 2019)

	Replication						
Variable	**Mean**	**Std Dev**	**10%**	**25%**	**Median**	**75%**	**90%**
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ESG$^{(PAIR)}$	0.44	0.218	0.166	0.278	0.417	0.586	0.76
ESG Uncertainty$^{(PAIR)}$	0.179	0.116	0.048	0.095	0.162	0.241	0.325
ESG$^{(ALL)}$	0.482	0.22	0.197	0.322	0.466	0.638	0.8
ESG Uncertainty$^{(ALL)}$	0.212	0.124	0.053	0.115	0.205	0.295	0.376

	Original						
Variable	**Mean**	**Std Dev**	**10%**	**25%**	**Median**	**75%**	**90%**
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ESG$^{(PAIR)}$	0.461	0.202	0.219	0.31	0.437	0.595	0.753
ESG Uncertainty$^{(PAIR)}$	0.18	0.112	0.051	0.097	0.162	0.246	0.33
ESG$^{(ALL)}$	0.49	0.206	0.239	0.337	0.472	0.63	0.788
ESG Uncertainty$^{(ALL)}$	0.207	0.124	0.051	0.11	0.195	0.291	0.373

	Replication - Original						
Variable	**Mean**	**Std Dev**	**10%**	**25%**	**Median**	**75%**	**90%**
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ESG$^{(PAIR)}$	-0.02	0.02	-0.05	-0.03	-0.02	-0.01	0.01
ESG Uncertainty$^{(PAIR)}$	0.00	0.00	0.00	0.00	0.00	-0.01	-0.01
ESG$^{(ALL)}$	-0.01	0.01	-0.04	-0.02	-0.01	0.01	0.01
ESG Uncertainty$^{(ALL)}$	0.01	0.00	0.00	0.01	0.01	0.00	0.00

This table reports the summary statistics of ESG Rating Uncertainty and ESG Rating measures for our replication of Avramov et al. (2022). Panel A presents pairwise ESG Rating Uncertainty across data providers (Replication, Original, and the difference R – O). Panel B shows pairwise correlations of ESG ratings. Panel C reports the quantile distribution of stock-level ESG characteristics for our replication, the original sample, and the differences. Reported statistics include mean, standard deviation, and selected quantiles (10%, 25%, median, 75%, 90%).

Online Appendix B. "Dissecting green returns" by Pástor et al. (*Journal of Financial Economics*, 2022) – Replication details

Online Appendix B.1. Sample

We utilize the same sample as Pástor et al. (2022) which consists of U.S. common stocks with the share codes of 10 or 11 from CRSP. Furthermore, we also use the same Environmental Pillar score and Environmental Pillar weights from MSCI. Pástor et al.'s (2022) study begins in November 2012 and ends in December 2020. We conduct our analysis over the same time period, and over an extended period that ends in June 2024 (considering our MSCI data availability).

Online Appendix B.2. Method

Following Pástor et al. (2022), we calculate their *unadjusted greenness score* by computing the distance between the highest-achievable score (10) and a firm's Environmental score, then multiplying it by the Environmental Pillar weight. We flip the sign to change the interpretation so that a higher number reflects a higher degree of greenness. Next, we center the greenness of each stock by subtracting the value-weighted market average for that month. Therefore, a positive (negative) score can be interpreted as being greener (browner) than the market. The top third of stocks is sorted into the green portfolio and the bottom third is sorted into brown portfolio. Both portfolios are value-weighted.

Online Appendix B.3. Single Specification Replication Results

Online Appendix Table B.1 shows that our replication over the period ending in December 2020 (as in the original study) reproduces the significant returns and alphas across five models, namely, Return (model with constant), CAPM, FF3, C4, and FF5. Online Appendix Figure B.1 corroborates this result: the replicated green and brown cumulative-return series closely track the published paths over the period November 2012–December 2020, yielding an almost identical GMB spread. However, extending our sample to June 2024 results in markedly smaller alphas, and except for model FF5, all alpha estimates are statistically insignificant.

Online Appendix Table B.1. Summary statistics

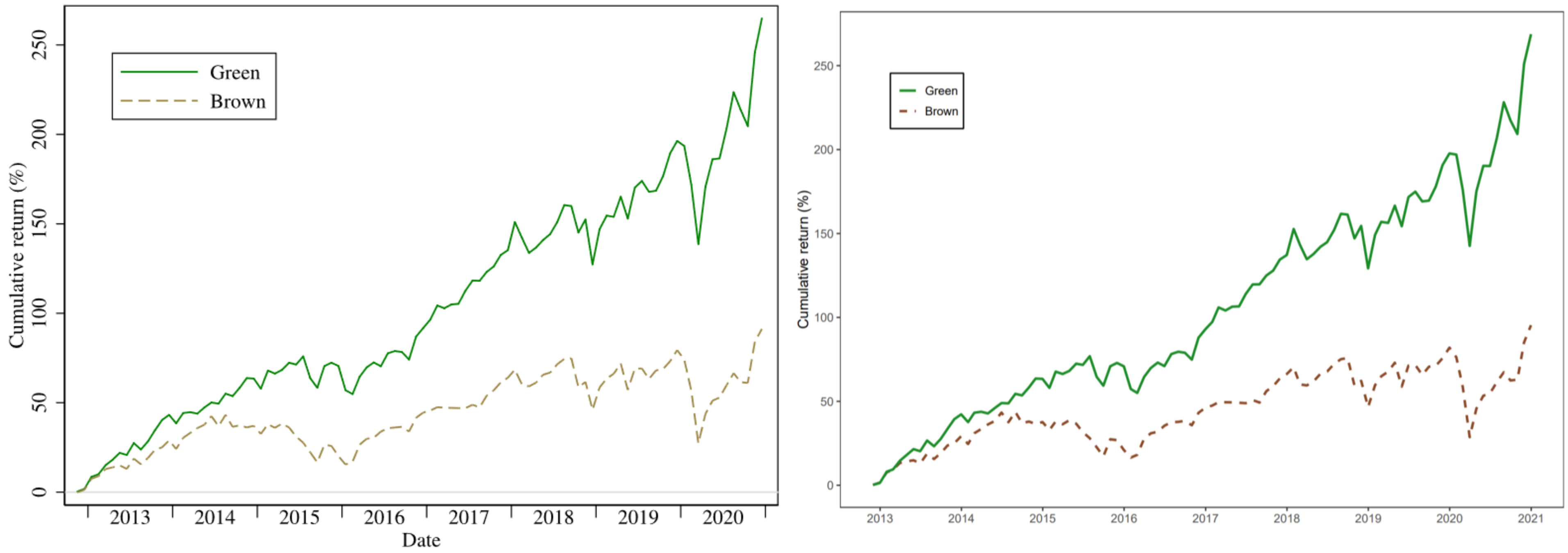
(Original study Table 3: GMB Performance (Page 411)**)**

		Return	**CAPM**	**FF3**	**C4**	**FF5**
		(1)	(2)	(3)	(4)	(6)
Original study	Return/Alpha	0.65	0.71	0.50	0.47	0.50
Nov 2012-Dec 2020	T-Value	**3.23**	**2.91**	**2.23**	**2.14**	**2.38**
Replication	Return/Alpha	0.66	0.73	0.52	0.50	0.51
Nov 2012-Dec 2020	T-Value	**3.76**	**3.27**	**3.14**	**3.04**	**3.25**
Replication	Return/Alpha	0.38	0.38	0.30	0.29	0.39
Nov 2012-Jun 2024	T-Value	1.47	1.27	1.69	1.75	**2.64**

This table reports return/alpha estimates (monthly, in percentage points) and T-statistics (in bold if significant) for the GMB spread from regressions using five models corresponding to columns (1), (2), (3), (4), and (6) in Pástor et al. (2022) Table 3: Return (model with constant), CAPM, Fama–French three-factor model (FF3), Carhart four-factor model (C4), and Fama–French five-factor model (FF5). We estimate all models with the Newey–West standard errors and report robust T-values. The original study has MSCI ENV rating data to March 2020 and extends MSCI rating data beyond March 2020 via a 12-month lookback. We use rating data through December 2020 and June 2024 to carry out our replications.

Overall, our estimates are robust for the original study period ending in December 2020. However, only the Fama-French five-factor (FF5) model alpha estimate remains significant for the extended replication ending in June 2024.

Online Appendix Figure B.1. Reported and replicated returns on value-weighted Green and Brown portfolios
(Original study Figure 3: Returns on value-weighted green and brown portfolios (Page 410))



Panel A. Pástor et al. (2022), Figure 3 (November 2012 – December 2020)

Panel B. Replication of Figure 3 (November 2012 – December 2020)

Explanation: This Appendix compares the reported returns from Pástor et al. (2022) on value-weighted green and brown portfolios (their Figure 3, p. 410, our Panel A) with our replication over the same period, November 2012–December 2020 (our Panel B). Pástor et al. report cumulative returns of 264.9% (green) and 91.3% (brown), implying a cumulative return difference (CRD) of 173.6% and a monthly Green–Minus–Brown (GMB) spread of 65 basis points (bps). Our replication yields closely aligned results: cumulative returns of 273% (green) and 93.1% (brown), a CRD of 180.1%, and a GMB of 66.1 bps. Using the publicly available portfolio returns from the replication package by Taylor, an author of this study, (https://data.mendeley.com/datasets/dnskbdnmsz/1), we further verify the robustness of the replication, with correlations of 0.9995 (brown) and 0.9993 (green).

Interpretation: Our single-specification replication virtually mirrors Pástor et al.'s (2022) over the same study period ending in December 2020. The Green and Brown return paths are almost identical.

Online Appendix C. "ESG Rating Disagreement and Stock Returns" by Gibson-Brandon et al. (*Financial Analysts Journal*, 2021) – Replication details

Online Appendix C.1. Sample

Gibson-Brandon et al.'s (2021) sample consists of S&P 500 stocks between 2010 and 2017.[51] They utilize ESG, ENV, SOC and GOV ratings from seven providers, namely, Refinitiv (REF), Bloomberg (BB), Sustainalytics (SUS), MSCI IVA (MSCI), MSCI KLD, FTSE, and Inrate. Considering our rating data availability, we conduct a replication of the relevant analyses using ratings assigned by four providers namely REF, BB, SUS,[52] and MSCI. Gibson-Brandon et al. (2021) identify these four as the most important rating providers (p. 107).[53]

Online Appendix Table C.1 compares the original study's and our rating coverage categorized by providers and rating types. Overall, our rating sample closely replicates Gibson-Brandon et al. (2021).

Online Appendix Table C.1. Rating data coverage (January 2010 – December 2017)
(Original study Table 1. ESG Data providers, p. 108)

Provider	**Original**	**Rep: ESG**	**Rep: ENV**	**Rep: SOC**	**Rep: GOV**
(1)	(2)	(3)	(4)	(5)	(6)
REF	438	479	479	479	479
SUS	459	450	449	449	449
BB	463	479	485	485	482
MSCI	456	452	452	452	452

This table compares the number of S&P 500 firms with available sustainability ratings (columns 3-6; ESG, ENV, SOC, GOV) assigned by one of the four examined providers (REF, SUS, BB, MSCI) in our replication sample and the original study's sample (column 2).

Online Appendix C.2. Method

Analogous to Gibson-Brandon et al. (2021), each month we compute percentile ranks for each rating and then calculate the standard deviation of all percentile ranks for a firm. Online Appendix Table C.2 compares the original study's and replicated descriptive statistics across providers and rating dimensions. Our results closely match those of Gibson-Brandon et al. (2021), with only minor deviations in the number of observations, while means and standard deviations are virtually identical. Online Appendix Table C.3 reports pairwise rating correlations across providers. Except for some differences in the GOV dimension, we obtain values which are quite close to those in Gibson-Brandon et al. (2021).

[51] We thank the authors, specifically Peter Schmidt, for advising us on their sample construction.

[52] We are grateful to Morningstar Sustainalytics for helping us obtain complete rating data of S&P 500 constituents.

[53] Gibson-Brandon et al. (2021, p. 107) cited this from a survey conducted by Wong et al. (2019).

Online Appendix Table C.2. Descriptive statistics (January 2010 – December 2017)

(Original study Table 2. Descriptive Statistics and Correlations, January 2010-December 2017, p. 109)

Panel A: ESG

	No of Observations		**Mean**		**StdDev**	
Provider (1)	**Original** (2)	**Replication** (3)	**Original** (4)	**Replication** (5)	**Original** (6)	**Replication** (7)
REF	42,087	45,862	0.501	0.5	0.289	0.289
SUS	44,078	43,030	0.501	0.5	0.289	0.289
BB	44,464	45,505	0.501	0.5	0.289	0.289
MSCI	43,775	42,893	0.501	0.5	0.289	0.289

Panel B: ENV

	No of Observations		**Mean**		**StdDev**	
Provider (1)	**Original** (2)	**Replication** (3)	**Original** (4)	**Replication** (5)	**Original** (6)	**Replication** (7)
REF	42,019	45,862	0.501	0.5	0.289	0.289
SUS	44,020	42,967	0.501	0.5	0.289	0.289
BB	37,624	46,069	0.501	0.5	0.289	0.287
MSCI	43,580	42,893	0.501	0.5	0.289	0.289

Panel C: SOC

	No of Observations		**Mean**		**StdDev**	
Provider (1)	**Original** (2)	**Replication** (3)	**Original** (4)	**Replication** (5)	**Original** (6)	**Replication** (7)
REF	42,087	45,862	0.501	0.5	0.289	0.289
SUS	44,078	42,967	0.501	0.5	0.289	0.289
BB	44,364	46,069	0.501	0.5	0.288	0.289
MSCI	43,775	42,893	0.501	0.5	0.289	0.289

Panel D: GOV

	Observations		**Mean**		**StdDev**	
Provider (1)	**Original** (2)	**Replication** (3)	**Original** (4)	**Replication** (5)	**Original** (6)	**Replication** (7)
REF	42,087	45,862	0.501	0.5	0.289	0.289
SUS	44,078	42,967	0.501	0.5	0.289	0.289
BB	44,464	45,781	0.501	0.5	0.282	0.275
MSCI	43,775	42,893	0.501	0.5	0.289	0.288

This table reports descriptive statistics of the four rating types ESG, ENV, SOC and GOV (Panels A–D respectively) across four examined rating providers (REF, SUS, BB, and MSCI) in the original sample of Gibson-Brandon et al. (2021) and our replication. For each provider and rating dimension, we report the number of firm-month observations (columns 2–3), the mean (columns 4–5), and the standard deviation (columns 6–7). The replicated values closely align with the original sample's values.

Online Appendix Table C.3. Pairwise rating correlations (Jan 2010 – Dec 2017)

(Original study Table 2. Descriptive Statistics and Correlations, January 2010-December 2017, p. 109)

Pair **(1)**	**ESG**		**ENV**		**SOC**		**GOV**	
	Original (2)	Rep (3)	Original (4)	Rep (5)	Original (6)	Rep (7)	Original (8)	Rep (9)
REF-SUS	0.752	0.632	0.706	0.602	0.617	0.480	0.331	0.267
REF-BB	0.750	0.713	0.647	0.719	0.685	0.567	0.432	0.204
REF-MSCI	0.396	0.373	0.233	0.276	0.266	0.231	0.132	0.135
SUS-BB	0.693	0.625	0.557	0.594	0.527	0.489	0.327	0.168
SUS-MSCI	0.434	0.419	0.357	0.367	0.303	0.309	0.135	0.148
BB-MSCI	0.303	0.297	0.187	0.193	0.202	0.186	0.060	-0.006

This table reports pairwise correlations of ratings across the four examined providers in the original sample of Gibson-Brandon et al. (2021) and our replication. For each rater pair, we present results for four rating dimensions (columns 2–9): ESG, ENV, SOC, and GOV. Each entry shows the correlation between two providers (column 1), with original values in even-numbered columns and replication values in odd-numbered columns. Overall, the replicated correlations quite closely track the original, though some differences emerge in the GOV dimension.

We create a monthly industry-adjusted rating disagreement score for each firm by subtracting the monthly industry mean (classified by the Fama-French 12-industry classification). Online Appendix Table C.4 compares the distributional properties of disagreement scores for firm-months in our sample and the corrected statistics provided by the authors.[54] As shown below, our replication values (mean, median, std) quite closely match the corrected figures, though there are some differences in higher-order moments such as skewness and kurtosis.

Online Appendix Table C.4. Summary statistics of rating Disagreement scores (January 2010 – December 2017)

(Original study Table A2. Summary statistics; page 122)

Rating (1)	**Sample** (2)	**N** (3)	**Mean** (4)	**StD.** (5)	**Min** (6)	**Max** (7)	**Median** (8)	**Skew** (9)	**Kurt** (10)
ESG	Original	45,405	0.195	0.078	0.008	0.471	0.191	0.251	-0.344
	Replication	45,078	0.181	0.092	0.002	0.513	0.169	0.584	-0.035
ENV	Original	45,103	0.199	0.075	0.009	0.473	0.198	0.113	-0.417
	Replication	45,075	0.191	0.092	0.007	0.553	0.181	0.485	-0.122
SOC	Original	45,405	0.220	0.078	0.005	0.468	0.219	0.049	-0.462
	Replication	45,075	0.207	0.097	0.004	0.544	0.197	0.395	-0.343
GOV	Original	45,405	0.240	0.077	0.010	0.472	0.241	-0.063	-0.356
	Replication	45,075	0.244	0.096	0.005	0.564	0.246	0.017	-0.502

This table reports summary statistics of rating disagreement scores for the four dimensions (ESG, ENV, SOC, GOV). The “original” values shown here were directly provided to us by the authors via personal communication. Reported statistics include the number of observations (column 3), mean (4), standard deviation (5), minimum (6), maximum (7), median (8), skewness (9), and kurtosis (10).

As in Gibson-Brandon et al. (2021), we sort stocks into quintiles from the lowest (Q1) to the highest (Q5) using industry-adjusted rating disagreement scores. We then form equally weighted portfolios which are rebalanced annually each January using December disagreement values.

[54] The authors provided us with the statistics after identifying an error in Table A.2 in their appendix (some observations included were before 2010). The authors confirmed that all other tables and results in their published study Gibson-Brandon et al. (2021) are not affected by this small error.

Online Appendix C.3. Results

Online Appendix Table C.5 reports the returns for Low (Q1), High (Q5), and the High–Low (Q5–Q1) long–short portfolio over the original study period 2010-2017 (see "Replication" row, Panels A-D) and the extended period 2010–2023 (see "Replication+" row, Panels A-D). For risk-adjusted performance, we compute Sharpe ratios (SR) and estimate alphas using the CAPM, the Fama–French Three-Factor model (FF3), the Carhart four-factor model (C4), and the Fama–French Five-Factor model (FF5) on the realized portfolio returns.

During 2010-2017, our replication aligns quite closely with the original study's H-L ESG and GOV rating disagreement portfolios. For ESG in Panel A (GOV in Panel D), we find similarly positive and statistically significant (insignificant) returns and alphas. There are some differences between our and their estimates for ENV and SOC. For ENV (Panel B), our estimates are likewise positive, though mostly insignificant. For SOC (Panel C), we report insignificant positive alphas while the original study documents insignificant negative values. Over the extended period 2010–2023, none of our estimates is significant.

Online Appendix Table C.5. Portfolio sorts on rating Disagreement scores (January 2010 – December 2017)

(Original study Table 5 Portfolio Sorts on Industry-Adjusted ESG Rating Disagreement, January 2010-December 2017 (T-statistics in parentheses), Panel A-D, page 116)

Panel A: ESG	**Portfolio**	**Return**	**N**	**StD.**	**SR**	**CAPM**	**FF3**	**C4**	**FF5**
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Original	Low Disp_adj	**1.124**	94	3.809	0.295	–0.088	–0.068	–0.055	–0.074
Replication	Low Disp_adj	**1.056**	93	3.786	0.279	–0.145	–0.121	–0.119	–0.139
Replication+	Low Disp_adj	**0.978**		4.755	0.206	–0.114	–0.063	–0.043	–0.106
Original	High Disp_adj	**1.336**	94	3.769	0.355	0.144	0.164	**0.211**	0.165
Replication	High Disp_adj	**1.276**	93	3.984	0.320	0.022	0.057	0.078	0.061
Replication+	High Disp_adj	**1.115**		5.112	0.218	–0.053	0.017	0.054	–0.008
Original	H-L Disp_adj	**0.212**		0.942	0.225	**0.232**	**0.232**	**0.267**	**0.239**
Replication	H-L Disp_adj	**0.219**		0.983	0.223	**0.167**	**0.178**	**0.197**	**0.200**
Replication+	H-L Disp_adj	0.137		1.067	0.129	0.061	0.080	0.097	0.097
	T-VALUES	**Return**	**N**	**StD.**	**SR**	**CAPM**	**FF3**	**C4**	**FF5**
Original	Low Disp_adj	**2.891**				–1.260	–0.97	–0.794	–1.049
Replication	Low Disp_adj	**2.733**				–1.761	–1.607	–1.600	–1.736
Replication+	Low Disp_adj	**2.666**				–0.982	–0.936	–0.604	–1.489
Original	High Disp_adj	**3.474**				1.561	1.792	**2.425**	1.611
Replication	High Disp_adj	**3.137**				0.243	0.709	0.985	0.924
Replication+	High Disp_adj	**2.828**				–0.340	0.208	0.624	–0.114
Original	H-L Disp_adj	**2.209**				**2.151**	**2.192**	**2.664**	**2.044**
Replication	H-L Disp_adj	**2.188**				**2.517**	**2.431**	**3.037**	**2.576**
Replication+	H-L Disp_adj	1.666				0.842	1.290	1.601	1.515

Online Appendix Table C.5. Portfolio sorts on rating Disagreement scores (January 2010 – December 2017) (continued)

Panel B: ENV	**Portfolio**	**Return**	**N**	**StD.**	**SR**	**CAPM**	**FF3**	**C4**	**FF5**
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Original	Low Disp_adj	**1.186**	93	4.057	0.292	–0.091	–0.043	–0.016	–0.036
Replication	Low Disp_adj	**1.075**	92	3.905	0.275	–0.156	–0.111	–0.085	–0.136
Replication+	Low Disp_adj	**0.981**		4.787	0.205	–0.112	–0.042	–0.016	–0.078
Original	High Disp_adj	**1.397**	93	3.906	0.358	0.163	**0.183**	**0.212**	**0.178**
Replication	High Disp_adj	**1.244**	92	3.957	0.314	–0.004	0.018	0.050	0.004
Replication+	High Disp_adj	**1.072**		4.961	0.216	–0.058	0.003	0.040	–0.032
Original	H–L Disp_adj	**0.211**		1.026	0.205	**0.254**	**0.226**	**0.228**	**0.213**
Replication	H–L Disp_adj	**0.169**		0.843	0.200	**0.152**	0.129	0.135	0.140
Replication+	H–L Disp_adj	0.091		0.974	0.093	0.055	0.046	0.055	0.046
	T–VALUES	**Return**	**N**	**StD.**	**SR**	**CAPM**	**FF3**	**C4**	**FF5**
Original	Low Disp_adj	**2.866**				–0.902	–0.50	–0.187	–0.455
Replication	Low Disp_adj	**2.698**				–1.329	–0.986	–0.726	–1.300
Replication+	Low Disp_adj	**2.657**				–0.770	–0.472	–0.171	–0.890
Original	High Disp_adj	**3.504**				1.848	**2.040**	**2.274**	**1.982**
Replication	High Disp_adj	**3.080**				–0.049	0.205	0.531	0.058
Replication+	High Disp_adj	**2.802**				–0.414	0.037	0.440	–0.413
Original	H–L Disp_adj	**2.010**				**2.482**	**2.226**	**2.322**	**2.005**
Replication	H–L Disp_adj	**1.960**				**2.196**	1.799	1.851	1.775
Replication+	H–L Disp_adj	1.211				0.989	0.869	1.060	0.861

Panel C: SOC	**Portfolio**	**Return**	**N**	**StD.**	**SR**	**CAPM**	**FF3**	**C4**	**FF5**
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Original	Low Disp_adj	**1.283**	93	3.841	0.334	0.071	0.121	0.139	0.116
Replication	Low Disp_adj	**1.095**	92	3.844	0.285	–0.120	–0.092	–0.089	–0.122
Replication+	Low Disp_adj	**0.985**		4.944	0.199	–0.141	–0.076	–0.044	–0.117
Original	High Disp_adj	**1.330**	94	3.996	0.333	0.067	0.096	0.134	0.081
Replication	High Disp_adj	**1.261**	93	3.937	0.320	0.017	0.038	0.059	0.028
Replication+	High Disp_adj	**1.101**		4.933	0.223	–0.033	0.021	0.043	–0.015
Original	H–L Disp_adj	0.047		0.965	0.049	–0.004	–0.024	–0.004	–0.035
Replication	H–L Disp_adj	0.166		0.913	0.181	0.138	0.130	0.148	0.150
Replication+	H–L Disp_adj	0.116		1.006	0.115	0.108	0.097	0.086	0.102
	T–VALUES	**Return**	**N**	**StD.**	**SR**	**CAPM**	**FF3**	**C4**	**FF5**
Original	Low Disp_adj	**3.273**				0.865	1.734	1.943	1.950
Replication	Low Disp_adj	**2.791**				–0.943	–0.801	–0.719	–1.034
Replication+	Low Disp_adj	**2.583**				–1.093	–0.993	–0.517	–1.444
Original	High Disp_adj	**3.262**				0.658	1.077	1.748	0.998
Replication	High Disp_adj	**3.137**				0.221	0.512	0.782	0.418
Replication+	High Disp_adj	**2.893**				–0.245	0.289	0.555	–0.217
Original	H–L Disp_adj	0.480				–0.058	–0.338	–0.073	–0.479
Replication	H–L Disp_adj	1.778				1.412	1.432	1.479	1.440
Replication+	H–L Disp_adj	1.490				1.714	1.593	1.322	1.598

Online Appendix Table C.5. Portfolio sorts on rating Disagreement scores (January 2010 – December 2017) (continued)

Panel D: GOV	Portfolio	Return	N	StD.	SR	CAPM	FF3	C4	FF5
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Original	Low Disp_adj	**1.246**	94	3.888	0.320	0.013	0.050	0.071	0.057
Replication	Low Disp_adj	**1.232**	93	3.874	0.318	0.006	0.024	0.022	0.014
Replication+	Low Disp_adj	**1.036**		4.728	0.219	–0.046	0.003	0.011	–0.038
Original	High Disp_adj	**1.284**	94	3.794	0.338	0.088	0.118	**0.167**	**0.120**
Replication	High Disp_adj	**1.324**	93	3.875	0.342	0.109	0.143	**0.180**	0.135
Replication+	High Disp_adj	**1.144**		4.917	0.233	0.026	0.097	**0.141**	0.081
Original	H–L Disp_adj	0.038		1.006	0.037	0.075	0.069	0.096	0.063
Replication	H–L Disp_adj	0.093		0.987	0.094	0.103	0.119	0.158	0.121
Replication+	H–L Disp_adj	0.108		1.129	0.096	0.072	0.093	0.130	0.120
	T–VALUES	**Return**	**N**	**StD.**	**SR**	**CAPM**	**FF3**	**C4**	**FF5**
Original	Low Disp_adj	**3.140**				0.129	0.535	0.777	0.702
Replication	Low Disp_adj	**3.115**				0.052	0.225	0.208	0.135
Replication+	Low Disp_adj	**2.839**				–0.356	0.039	0.130	–0.450
Original	High Disp_adj	**3.315**				1.320	1.871	**2.749**	**2.577**
Replication	High Disp_adj	**3.349**				1.212	1.691	**2.130**	1.635
Replication+	High Disp_adj	**3.015**				0.178	1.490	**1.978**	1.266
Original	H–L Disp_adj	0.366				0.742	0.680	0.954	0.696
Replication	H–L Disp_adj	0.919				1.135	1.292	1.529	1.074
Replication+	H–L Disp_adj	1.244				0.878	1.295	1.856	1.471

Explanation: This table reports performance of portfolio sorts on industry-adjusted rating disagreement scores for the four dimensions (Panels A–D: ESG, ENV, SOC, GOV) as reported in Gibson-Brandon et al. (2021) (see "Original" row), and obtained in our replication using their single-specification method over the original study period 2010-2017 (see "Replication" row) and the extended period 2010-2023 (see "Replication+" row). Reported values include monthly portfolio returns (Return, column 3), number of months (N, column 4), returns standard deviation (StD. column 5), Sharpe ratio (SR, column 6), and factor-model alphas from CAPM, FF3, C4, and FF5 models (columns 7–10), with T-statistics (in bold if significant) shown in the lower part of each Panel. Results are presented for Low Disagreement, High Disagreement, and High minus Low Disagreement portfolios.

Interpretation: Our H-L rating disagreement portfolio performance tracks the original results quite closely for ESG and GOV, and to some extent, ENV. We mainly differ in that we get insignificant positive results for SOC, while the original values are insignificant negative. In the extended period, our estimates are generally insignificant and of lower magnitude.

Online Appendix D. Main empirical analysis using our multi-specification method

Online Appendix Table D.1. Descriptive statistics of our normalized rating scores

Provider	Rating	Unique Firms	Yearly Obs	Mean	Median	Std Dev	Min	Max
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Refinitiv	ESG	3,931	28,194	40.1	36.1	20.6	0.0	100.0
Refinitiv	ENV	3,930	28,190	24.4	15.7	27.0	0.0	100.0
Refinitiv	SOC	3,930	28,190	40.8	37.1	21.7	0.0	100.0
Refinitiv	GOV	3,888	27,878	47.3	47.0	23.1	0.0	100.0
MSCI	ESG	3,622	24,894	43.6	42.0	19.9	0.0	100.0
MSCI	ENV	3,622	24,894	45.7	45.0	21.9	0.0	100.0
MSCI	SOC	3,622	24,894	43.2	42.0	15.7	0.0	100.0
MSCI	GOV	3,622	24,888	58.6	59.0	19.8	0.0	100.0
Bloomberg	ESG	526	3,835	43.7	43.5	21.0	0.0	100.0
Bloomberg	ENV	1,029	7,341	21.4	12.8	23.7	0.0	100.0
Bloomberg	SOC	1,029	7,341	24.2	18.2	19.2	0.0	100.0
Bloomberg	GOV	1,115	7,974	64.4	66.0	14.9	0.0	100.0
RobecoSAM	ESG	1,358	6,295	38.7	33.0	28.1	0.0	100.0
RobecoSAM	ENV	1,358	6,295	38.8	33.0	27.2	0.0	100.0
RobecoSAM	SOC	1,358	6,295	34.6	27.0	28.3	0.0	100.0
RobecoSAM	GOV	1,358	6,295	45.6	43.0	25.9	0.0	100.0
Sustainalytics	ESG	2,357	18,742	47.3	48.0	19.6	0.0	100.0
Sustainalytics	ENV	1,099	10,308	53.4	53.0	25.3	0.0	100.0
Sustainalytics	SOC	1,099	10,308	55.9	59.0	21.1	0.0	100.0
Sustainalytics	GOV	1,099	10,308	62.2	68.4	24.9	0.0	100.0
Agreement	ESG	1,870	12,334	47.2	45.6	16.0	5.1	93.8
Agreement	ENV	1,488	11,994	41.7	40.1	19.3	0.0	96.4
Agreement	SOC	1,488	11,994	44.2	43.5	15.0	3.4	93.4
Agreement	GOV	1,491	12,061	58.2	59.2	13.9	2.5	96.9
Disagreement	ESG	1,870	12,334	16.8	16.5	7.5	0.3	48.9
Disagreement	ENV	1,488	11,994	20.7	19.7	9.4	0.0	57.2
Disagreement	SOC	1,488	11,994	19.6	19.5	7.5	0.3	46.8
Disagreement	GOV	1,491	12,061	18.3	17.6	8.4	0.3	54.5

Explanation: This table features the statistics of normalized Environmental (ENV), Social (SOC), Governance (GOV), and ESG rating scores issued by Refinitiv (2002-2022), MSCI (2007-2022), RobecoSAM (2016-2022), Bloomberg (2015-2022), and Sustainalytics (2009-2022). We also utilize our self-constructed rating Agreement and Disagreement scores (2009-2022). Sustainalytics changed its rating method from best-in-class to a risk-based approach in 2019. Before 2019, a high Sustainalytics score indicated an asset with "good" ESG values. From 2019 onwards, a high Sustainalytics score indicates an asset with a high ESG risk. Therefore, we invert Sustainalytics's rating scores assigned during the period 2019-2022 (Inverted score=100 – score) to make them consistent with Sustainalytics ratings issued during the preceding period, and consistent with ratings issued by other providers. Thus, a score of 10 indicating low risk in the new system (2019-2022) would now get an inverted high score of 90 which represents a good ENV/SOC/GOV/ESG value.

Column (3) shows the unique number of firms with a sustainability rating score, column 4 includes the yearly firm observations. Columns (5-9) present the statistics of scores specific to each rating type and each provider.

Online Appendix Table D.2. Percentage of sample excluded by the decision made at each individual node in Figure 1.

Provider	Rating	Excl Finance	Excl. Utilities	Negative book value	Negative earnings	Market cap < $300 million	Price < $5
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Refinitiv	ESG	16.18	4.52	4.94	23.08	11.31	6.62
	ENV	16.17	4.52	4.94	23.08	11.31	6.62
	SOC	16.17	4.52	4.94	23.08	11.31	6.62
	GOV	16.29	4.55	4.95	22.93	11.31	6.55
MSCI	ESG	15.77	4.38	4.92	21.73	5.19	4.6
	ENV	15.77	4.38	4.92	21.73	5.19	4.6
	SOC	15.77	4.38	4.92	21.73	5.19	4.6
	GOV	15.77	4.38	4.92	21.73	5.19	4.6
Bloomberg	ESG	11.06	10.15	4.93	12.23	0.81	1.62
	ENV	13.61	5.87	6.16	17.15	1.39	2.24
	SOC	13.61	5.87	6.16	17.15	1.39	2.24
	GOV	14.34	5.77	5.86	16.56	1.34	2.18
RobecoSAM	ESG	14.6	6.1	5.21	14.63	2.93	2.52
	ENV	14.6	6.1	5.21	14.63	2.93	2.52
	SOC	14.6	6.1	5.21	14.63	2.93	2.52
	GOV	14.6	6.1	5.21	14.63	2.93	2.52
Sustainalytics	ESG	17.07	5.02	4.3	16.7	13.34	5.29
	ENV	15.35	6.68	4.39	11.32	0.81	1.27
	SOC	15.35	6.68	4.39	11.32	0.81	1.27
	GOV	15.35	6.68	4.39	11.32	0.81	1.27
Agreement	ESG	15.28	6.34	4.92	13.63	1.03	1.55
	ENV	14.88	6.4	5.13	13.86	1.09	1.64
	SOC	14.88	6.4	5.13	13.86	1.09	1.64
	GOV	14.98	6.38	5.14	13.87	1.09	1.65
Disagreement	ESG	15.28	6.34	4.92	13.63	1.03	1.55
	ENV	14.88	6.4	5.13	13.86	1.09	1.64
	SOC	14.88	6.4	5.13	13.86	1.09	1.64
	GOV	14.98	6.38	5.14	13.87	1.09	1.65

Explanation: This table reports, for each provider (Refinitiv, MSCI, Bloomberg, RobecoSAM, Sustainalytics) and for our self-calculated Agreement/Disagreement scores, the proportion of firm–month observations removed in each step of our portfolio construction outlined in Figure 1. Column (2) indicates the rating type (ESG, ENV, SOC, GOV). Columns (3)–(8) show the percentage excluded by the respective decision nodes: excluding Finance (financials sector), excluding Utilities (utilities sector), excluding negative book value, negative earnings, excluding Market Capitalization<$300 million, and excluding Price<$5. Percentages are calculated relative to the provider–rating universe and show the effect of each decision node on its own, irrespective of the other nodes; therefore, the numbers are not additive, and totals do not sum to 100%. Values are in percentage (two-decimal rounding).

Overall, decisions concerning excluding the Financials sector and excluding negative earnings eliminate the largest fractions in most universes. With a few exceptions, within a provider the excluded shares are *mostly* similar across ESG, ENV, SOC, GOV dimensions.