ESSAYS ON SUSTAINABLE EQUITY INVESTING

EMPIRICAL EVIDENCE FROM RATINGS, INDICES, AND PORTFOLIO OPTIMIZATION

Dissertation

zur Erlangung des Grades eines Doktors der Wirtschaftswissenschaft

der Rechts- und Wirtschaftswissenschaftlichen Fakultät

der Universität Bayreuth

Vorgelegt

von

Jan Josef Heldmann

aus

Friedenfels

Dekan:	Prof. Dr. Claas Christian Germelmann
Erstberichterstatter:	Prof. Dr. Klaus Schäfer
Zweitberichterstatter:	Prof. Dr. David Stadelmann
Tag der mündlichen Prüfung:	10. August 2026

Acknowledgements

Completing this dissertation would not have been possible without the support of many wonderful people who accompanied me on this journey.

First and foremost, I would like to express my sincere gratitude to Prof. Dr. Klaus Schäfer for giving me the opportunity to pursue this doctorate. His support and the encouraging environment he created made it possible for me to succeed in the first place. I am also grateful to my colleagues Alexander-Michael Gerk, Andreas Horn, Christine Landmann, Manuel Brinkmann, and Michael Bruischütz at the Chair of Banking and Finance at the University of Bayreuth. Because of you and Prof. Schäfer, I enjoy each day at the chair. I truly appreciate our collegial, supportive atmosphere, our academic and non-academic conversations, and the many happy moments we share. Special thanks go to my co-authors, Thomas Brückner and Manuel Brinkmann, and especially to Huong Dieu Dang for her guidance, support, and contributions, which were essential to this dissertation.

I want to sincerely thank my friends and extended family for reminding me that there is life beyond my studies. I am also deeply grateful to my parents, Margareta and Manfred, and my sister, Vera. Thank you for your unconditional support throughout my entire academic journey and for always believing in me. To my girlfriend, Annalena, thank you for being my anchor through it all, for your endless patience and support, for always standing behind me, for enduring countless late nights and for giving me the strength to keep going. I cannot thank you enough for your support and love.

Finally, I would like to dedicate this dissertation to the memory of my grandmother, Bella Oma. I wish you could have witnessed this moment. Your love, spirit and endurance continue to inspire me every day.

Abstract

Sustainable investing has become a significant segment of global capital markets and a popular research field in academia, where two fundamental questions are of particular importance. First, whether sustainable investment strategies affect financial performance, and second, whether they achieve their non-financial sustainability objectives. This dissertation addresses both questions through four empirical papers that examine different dimensions of sustainable equity investing across investment strategies and outcome measures.

The first two papers jointly evaluate 20 MSCI equity indices across four regions (World, U.S., Europe, and Emerging Markets), comparing four sustainable index variants, namely, Climate Change (CC), Paris-Aligned Benchmark (PAB), Socially Responsible Investment (SRI), and SRI Filtered PAB (SRI PAB), against their standard counterparts. On the financial side, all four sustainable variants outperform their standard counterparts, with a cumulative return differential ranging from 4.7% to 13.7% over 2015 to 2023. On the sustainability side, an investment in the sustainable variants instead of the standard index results in carbon emission reductions between 53% and 77% by the end of 2022. A regression framework that controls for sector, country development level, and size composition, as well as time-specific market events, confirms that both the financial outperformance and the emission reductions are statistically significant and linked to the index construction criteria in the period of 2015–2022. Additionally, these findings hold against a comprehensive battery of robustness tests. However, broader sustainability assessments through environmental and ESG ratings show that improvements depend on both the index variant and the rating provider. Among the variants, the PAB indices achieve the largest carbon emission reductions, while the SRI variants deliver the strongest improvements in broader sustainability scores.

The third paper shifts from evaluating existing index products to investigating what is technically achievable through active portfolio optimization. Using a self-built multi-factor risk model applied to the STOXX Europe 600 over 2011–2024, the paper constructs low-carbon portfolios that minimize Scope 1+2 carbon intensity subject to a tracking error budget and liquidity, sector, country, and factor-exposure constraints. The results demonstrate that carbon intensity reductions of 61% are achievable at a realized tracking error of only 0.64%, increasing to over 93% at a tracking error budget of 2%. Further, there is no evidence of a financial performance penalty, as information ratios range from about 0.5 to 0.8. An attribution analysis shows that within-sector stock selection, rather than sector tilts, accounts for the dominant share of both carbon reductions and active returns.

Since sustainability ratings serve as an important input for portfolio construction and sustainable investment decisions, the fourth paper examines whether such ratings from five major providers (Refinitiv, MSCI, Bloomberg, RobecoSAM, and Sustainalytics), as well as self-constructed provider agreement and disagreement ratings, constitute a robust investment and pricing signal. Using a multi-specification framework that systematically varies portfolio construction choices across multiple decision nodes for U.S. equities over 2003–2024, the paper finds no robust evidence that higher-rated portfolios or rating-based factors generate superior risk-adjusted returns. Similarly, adding a sustainability rating factor to the Fama-French six-factor model yields no meaningful improvement in explaining the cross-section of returns. Additionally, the replications of three influential studies show that their findings regarding sustainability ratings become fragile under alternative specifications and extended sample periods.

The findings of this dissertation suggest that sustainable equity strategies, whether implemented through commercial index products or active portfolio optimization, do not impose a systematic performance penalty and, if anything, tend to deliver modest outperformance in many cases, while achieving measurable sustainability improvements. This suggests that financial and sustainability objectives are not inherently in conflict, at least within the equity asset class and the markets and time periods examined. However, sustainability ratings are only weakly informative for equity portfolio performance and factor models, and both financial and sustainability outcomes are sensitive to methodological choices, including the rating provider and portfolio construction decisions.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ACRD	Average Cumulative Return Differential
AuM	Assets under Management
CAPM	Capital Asset Pricing Model
CC	Climate Change
CDP	Carbon Disclosure Project
CRD	Cumulative Return Differential
CTB	Climate Transition Benchmark
CWD	Cumulative Weighting Difference
EM	Emerging Markets
EMH	Efficient Market Hypothesis
ENV	Environmental
ESG	Environmental, Social, and Governance
ETF	Exchange-Traded Fund
EU	European Union
EVIC	Enterprise Value Including Cash
FF6	Fama-French Six-Factor Model
GCE	Green Criteria Effectiveness
GMB	Green-Minus-Brown
GOV	Governance
GRS	Gibbons, Ross, and Shanken Test
HDI	Human Development Index
MPT	Modern Portfolio Theory
OLS	Ordinary Least Squares
PAB	Paris-Aligned Benchmark
SOC	Social
SRI	Socially Responsible Investment/Investing
SRI PAB	SRI Filtered PAB
TE	Tracking Error
UN	United Nations

1 Introduction

1.1 Motivation and Background

Sustainable investing has become a significant segment of global capital markets. As of September 2025, assets in sustainable funds reached a new record high of $3.7 trillion (Bioy et al., 2025). Among these, climate-themed funds in particular saw a substantial increase from about $30 to $652 billion between 2018 and 2025 (Lee et al., 2026), as shown in Figure 1.

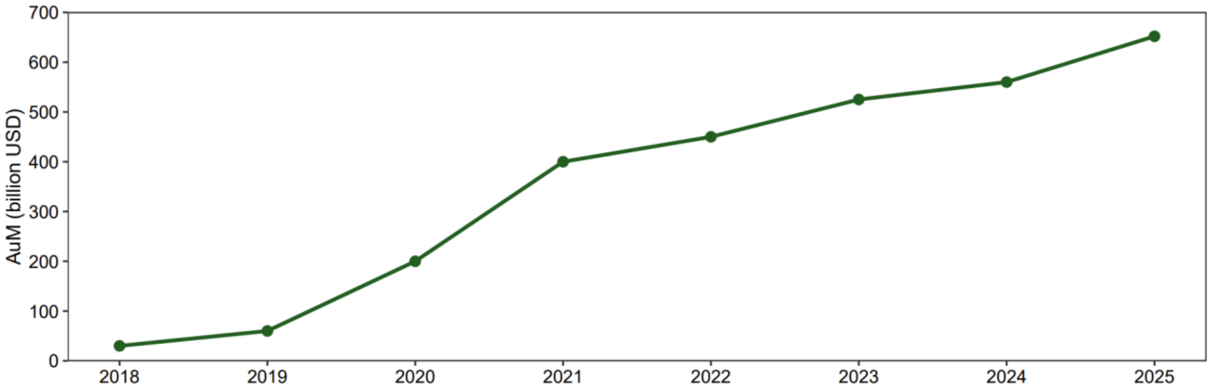


Figure 1. Growth of AuM in climate funds

Notes: This figure shows global assets under management (AuM) in climate-themed funds from 2018 to 2025. Values are reported in billion USD.
Source: Lee et al. (2026).

This growth reflects broad investor demand. According to a recent global survey, 88% of individual investors report an interest in sustainable investing and more than half plan to increase their sustainable allocations (Morgan Stanley Institute for Sustainable Investing, 2025). At the same time, the use of environmental, social, and governance (ESG) information in sustainable investment decisions has continued to grow. Nine out of ten institutional investors reported an increased use over the past year (Bell and Taylor, 2024) and 81% of asset managers stated that the importance of ESG considerations has further intensified (Index Industry Association, 2023).[1]

The motives behind sustainable investing are both financial and non-financial. Among retail investors, non-financial motives include ethical considerations as well as prosocial and environmental preferences (Giglio et al., 2025; Siemroth and Hornuf, 2023; Gutsche et al., 2023; Heeb et al., 2023; Riedl and Smeets, 2017).[2] Institutional investors and asset managers integrate sustainability considerations for similar reasons, but also face growing regulatory, reputational, and competitive pressure (Kolasa and Sautner, 2025; Krueger et al., 2020; Otchere

[1] However, in the same survey, 66% of investors also reported that their institution is likely to scale back ESG considerations in allocation decisions (Bell and Taylor, 2024).
[2] The relative importance of these motives varies across studies. A substantial share (48%) of retail investors also reports no sustainability-related investment motive at all (Giglio et al., 2025).

et al., 2025; Balp and Strampelli, 2022). Aside from these considerations, both groups share an interest in reducing their exposure to climate-related risks, including transition risks, physical climate risks, and litigation risks (Bell and Taylor, 2024; Setzer and Higham, 2025; FTSE Russell, 2022).

Even though sustainable investing extends beyond climate change mitigation, climate-related risks have become a central focus, with 46% of investors identifying climate change as a key factor in their investment policy (Robeco, 2025). This trend is supported by recent financially material developments, as global average temperatures have breached the 1.5°C threshold above pre-industrial levels (World Meteorological Organization, 2025; Copernicus Climate Change Service, 2025) and financial losses from natural catastrophes have risen sharply (Munich Re, 2025; Swiss Re, 2024).

Relatedly, sustainable investing also has become a prominent topic in academic research. Bibliometric analyses show that the number of publications related to sustainable investing has grown substantially over the past three decades (Chalissery et al., 2023; Luo et al., 2022), especially after 2020 (Gazali et al., 2025). Additionally, peer-reviewed journals dedicated to this specific topic have emerged, such as the Journal of Sustainable Finance & Investment or the Journal of Climate Finance. The overall volume of research has also led to large meta-analyses that try to synthesize results from thousands of individual papers (Atz et al., 2023; Whelan et al., 2021; Friede et al., 2015).

Naturally, two fundamental questions arise in the context of this growing relevance of sustainable investing. First, how does sustainable investing affect financial performance? And second, do sustainable investment strategies achieve their non-financial objectives? Section 1.3 reviews the current state of the literature on these questions and shows that both remain insufficiently resolved despite the large body of existing research. Section 1.4 then presents how this dissertation contributes to addressing these two questions. Before, however, the following section describes what sustainable investing encompasses, as there is no single, universally agreed-upon definition.

1.2 Characterizing Sustainable Investing

Sustainable investing is a subfield of sustainable finance, which also includes, among others, sustainable lending and the management of ESG-related risks in banking and insurance (Schoenmaker and Schramade, 2019). Within this broader field, sustainable investing focuses on how sustainability considerations influence investment decisions in capital markets.

Specifically, in this dissertation, the analysis is restricted to the equity side of sustainable investing. Since sustainable investing itself comprises a diverse set of concepts, it is helpful to distinguish between the different labels sustainable investing covers and the strategies used to implement them.

The label sustainable investing is commonly used as an umbrella term which includes a wide range of related and often overlapping concepts with different emphases, such as environmental, social, and governance (ESG) investing, socially responsible investing (SRI), ethical investing, as well as green or low-carbon investing.

ESG investing describes the incorporation of environmental, social, and governance factors, also called pillars, into the investment analysis. Each pillar consists of various dimensions, for example, emissions and resource use (environmental), workforce practices and human rights (social), and management structure and shareholders (governance).[3] To quantify ESG, rating providers aggregate these dimensions into pillar scores and weight these to create a single ESG score, although these can vary considerably across providers because of differences in methodologies (Berg et al., 2022).

Socially responsible investing (SRI) generally emphasizes values-based investing for social, environmental, or moral reasons. This includes sector-specific (alcohol, tobacco, gambling) or environmental (fossil fuel, nuclear energy) concerns, ethical issues (animal testing), faith-based criteria (abortion, dietary restrictions), or the promotion of social good.[4] Many SRI investors also integrate ESG scores into their decision-making process (Sherwood and Pollard, 2023, Chapter 2). Ethical investing broadly overlaps with SRI but tends to have a stronger emphasis on moral convictions as the primary driver of investment decisions.

Green investing and low-carbon investing focus on environmental objectives, especially climate change mitigation. They target reduced carbon emissions or align portfolios with specific climate goals such as those established by the Paris Agreement. In the European Union (EU), the European Commission (2020) also defined two regulated climate benchmark labels, the EU Climate Transition Benchmark (CTB) and the EU Paris-Aligned Benchmark (PAB).[5]

[3] Since ESG providers use different methodologies, terminology within pillars may vary slightly across data sources (Berg et al., 2022). The examples listed here follow the category structure of LSEG (2024a).
[4] These considerations can be subjective. A commonly excluded sector for ethical reasons is weapons manufacturing. However, after the Russian invasion of Ukraine in 2022 and the so-called *Zeitenwende*, a renewed debate emerged about this classification as many emphasized the importance of defense spending for protecting democratic values and national security.
[5] Both CTB and PAB require an annual decarbonization of 7% compared to their conventional benchmark. They differ mainly in their initial emission reduction levels which are 30% for the CTB and 50% for the PAB.

These labels are not mutually exclusive. A single investment product might be described as an ESG fund, a sustainable fund, and a responsible fund simultaneously. The description might also vary across regions and institutions, and change over time.

While the labels above describe what investors care about, strategies describe how these motives and concerns are implemented into portfolios. While various frameworks exist for categorizing sustainable investment strategies (Sherwood and Pollard, 2023, Chapter 3; Global Sustainable Investment Alliance, 2025; European Sustainable Investment Forum, 2016), the methods they describe are very similar, differing mostly in naming conventions and grouping rather than in the methods themselves. Drawing on these frameworks, four common strategy groups are distinguished for the purpose of this characterization: exclusionary, selective, integrative, and engagement.

Exclusionary strategies focus on filtering certain stocks or sectors from the investment universe. The two most common approaches are product-based screening, which removes companies involved in unwanted business activities (e.g., tobacco production), and norms-based screening, which excludes companies based on behavioral grounds (violation of international standards or norms, e.g., United Nations (UN) Global Compact[6]). These strategies are the oldest form[7] of sustainable investing and comparatively simple to implement, usually at the very beginning of the portfolio construction process with an exclusion list.

Selective strategies focus on favoring stocks with (relatively) strong sustainability characteristics. These characteristics are measured through ESG scores, emissions metrics, or variations thereof. Selective strategies differ in how they determine which stocks qualify for inclusion. Positive screening sets a minimum threshold for the relevant characteristic, for example, only including companies with an ESG score above a certain level. Best-in-class selection takes a relative approach and includes the top-performing companies, either across the entire market or within each sector. Thematic investing represents a specific form of selection that targets companies in sustainability-related fields, such as renewable energy, clean technology, or sustainable agriculture.

Integrative strategies incorporate sustainability factors into the investment process without necessarily excluding or selecting specific stocks. ESG integration treats environmental, social, and governance information as additional inputs into the portfolio construction process, often

[6] The UN Global Compact lists ten principles for corporate sustainability in four areas: human rights, labor, environment, and anti-corruption. See https://unglobalcompact.org/what-is-gc/mission/principles.

[7] Although there is no definitive starting point of sustainable investing, some of its roots can be traced back to the 1700s, where faith-based investors like the Quakers would not invest into war or slavery (Schueth, 2003).

with the idea that sustainability factors incorporate risks which may be overlooked by a traditional analysis. For instance, this can result in a portfolio with the same number of constituents as its benchmark, yet with more favorable sustainability characteristics as portfolio weights are shifted away from firms with weaker sustainability profiles toward those with stronger ones.

Engagement strategies differ from the previous three approaches, which focus on portfolio construction, insofar as they influence corporate sustainability practices directly and thus try to change corporate behavior from within. Owning shares of a company allows investors to vote on shareholder resolutions, file shareholder proposals, or engage directly with the company management on sustainability issues (Dimson et al., 2015). This form of engagement is called active ownership. A related approach is impact investing, which explicitly targets quantifiable social or environmental results besides financial returns (Global Impact Investing Network, 2025). However, this dissertation focuses on exclusionary, selective, and integrative strategies.

Like the labels, these strategies are not mutually exclusive. Many sustainable investment products apply multiple approaches by layering strategies to achieve their objectives. The MSCI SRI Filtered Paris-Aligned Benchmark Index, one of the indices that is examined in this dissertation, illustrates this well (MSCI, 2024a). The index combines socially responsible investing principles with Paris Agreement alignment, as reflected by its name. It first applies extensive exclusionary screening across controversial activities, then employs a selective best-in-class approach by including only the top 25% of ESG-rated companies within each sector and thereafter integrates climate objectives through the Paris-Aligned Benchmark overlay requiring a 50% reduction in weighted average carbon intensity and a 7% annual decarbonization trajectory. This layered approach results in an index that excludes approximately 75–90% of the parent index constituents.[8]

The breadth of labels and strategies described above shows that sustainable investing is not a single category or investment product but a diverse collection of approaches with different objectives and methods. Accordingly, the question of whether sustainable investing delivers on its promises is complex and multi-faceted. The following section reviews this debate.

[8] A detailed description of the index methodologies is provided in Papers 1 and 2.

1.3 Financial Performance and Sustainability Outcomes

The debate revolves around two important questions sustainable investors face. First, how does the fact that the investment is sustainable affect its financial performance? Second, does sustainable investing actually achieve favorable sustainable outcomes?

1.3.1 Financial Performance

The first question can be examined from a theoretical and empirical perspective. Traditional financial theory and models like the Modern Portfolio Theory (MPT) and the Capital Asset Pricing Model (CAPM) imply that sustainable investing could potentially lead to suboptimal investment results. The MPT by Markowitz (1952, 1959) states that, for a given expected return, the combination of not perfectly correlated assets can reduce portfolio variance. The set of assets that achieve the minimum variance for each level of expected return is called the efficient frontier. The core idea of the CAPM, developed by Sharpe (1964), Lintner (1965), and Mossin (1966), is that each stock carries two kinds of risk[9], the idiosyncratic risk specific to the company and the systematic risk which is its exposure to the market. While the systematic risk from the market cannot be eliminated, idiosyncratic risk can be reduced through diversification, which is achieved by increasing the number of assets in the portfolio. Since sustainable investment strategies, like negative screening for example, decrease the number of possible assets by excluding them from the investable universe, investors may not be able to find the optimal (unconstrained) mean-variance efficient frontier or the desired level of diversification for their portfolio.

The Efficient Market Hypothesis (EMH) suggests that in informationally efficient markets, equity prices fully reflect all available information (Fama, 1970). If this is the case, the choice between sustainable and conventional investing should not systematically affect expected financial returns. Any positive, negative, or negligible effects of sustainability characteristics on firm fundamentals and investor expectations are already incorporated into the current stock price through the actions of rational investors. Therefore, restricting the investment universe to sustainable stocks or tilting portfolio weights based on sustainable criteria would not generate superior expected returns, as the market has already priced in these factors.

More recent studies, such as those by Pedersen et al. (2021) and Pástor et al. (2021), propose theoretical frameworks that extend traditional theories by including investors' ESG preferences and the pricing implications of sustainability characteristics. Pedersen et al. (2021)

[9] Later models show that expected returns depend on multiple systematic risk factors rather than market risk alone (Fama and French, 1993; Carhart, 1997).

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. Pástor et al. (2021) argue that the outperformance of assets with good ESG metrics hinges on the taste of investors and the need to hedge climate risks. 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. 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. These theoretical frameworks allow positive, negative, or neutral sustainable investment performance depending on market conditions, investor preferences, and the informational value of ESG characteristics.

Besides theoretical approaches, there have been numerous empirical studies conducted in the field of sustainable investing over the last two decades with several meta-studies attempting to synthesize their results. These meta-studies usually cover both accounting-based financial performance, for example, profitability, growth, cash flows, or valuation measures such as Tobin's Q, and market-based financial performance, for example, stock or portfolio returns, risk-adjusted returns, or alpha. Since this dissertation examines sustainable investment strategies and their outcomes at the portfolio level, the focus is on the market-based financial performance (hereafter just financial performance) of these meta-studies.

An early meta-study conducted by Margolis et al. (2009) analyzes 251 studies and finds a small but positive association between corporate social performance and financial performance. Friede et al. (2015) aggregate evidence from more than 2,000 empirical studies of which about 290 analyze the relationship between equities and ESG. Of these, 52.2% find a positive relationship and only 4.4% a negative one. The most recent comprehensive meta-study by Atz et al. (2023) analyzes 1,141 papers and 27 meta-reviews covering the period from 2015 to 2020. From these, equity papers (n=80) report 38% positive, 49% neutral/mixed, and 14% negative relationships with financial performance. Atz et al. (2023) further differentiate financial performance by their strategy. The results are most favorable for integrative strategies (ESG integration), largely neutral/mixed for selective strategies (pooled strategies), and more split for exclusionary strategies (negative screening and divesting). Table 1 reports the corresponding shares. Atz et al. (2023) additionally run regressions which show that integrative strategies are significantly positively associated with financial performance, while exclusionary and selective strategies have insignificantly negative coefficients.

Table 1. Sustainability-performance relationship in literature

Asset class/Strategy	Studies	Positive	Neutral/Mixed	Negative
Equities	80	38%	49%	14%
Fixed income	24	33%	54%	12%
Negative screening & divesting	13	31%	38%	31%
Pooled strategies	42	21%	62%	17%
ESG integration	34	59%	38%	3%

Notes: Percentages report the share of investor studies that find a positive, neutral/mixed, or negative relationship between sustainability and financial performance. Categories are not mutually exclusive, as a study can receive multiple codes or an “other/not applicable” code. In this dissertation’s taxonomy, negative screening & divesting corresponds to exclusionary strategies, pooled strategies corresponds to selective strategies, and ESG integration corresponds to integrative strategies.
Source: Based on Table 1, Panel B, in Atz et al. (2023).

However, the overall results still demonstrate that even within strategies, substantial heterogeneity exists. This variation reflects systematic differences in research design, including geographic coverages, sample periods, ESG rating providers, and portfolio construction periods. Berg et al. (2022) document that ESG ratings differ substantially across providers, with measurement differences (56%) and scope (38%) accounting for the majority of variance in scores. More broadly, recent research also highlights that methodological choices in portfolio construction can materially alter empirical results (Walter et al., 2024; Soebhag et al., 2024; Beyer and Bauckloh, 2024; Henriquez-Salman, 2025). This sensitivity to specification choices may explain part of why decades of research have produced such a variation in answers to the question of financial performance. Section 1.4 will discuss how this dissertation addresses specific gaps in this contested literature.

Yet, financial performance represents only one dimension of evaluating sustainable investing. As discussed in Section 1.1, many investors also have non-financial motivations for investing sustainably. This raises the second question, whether sustainable strategies actually deliver on these sustainability objectives, which will be answered in the following section.

1.3.2 Sustainability Outcomes

A preliminary question is why sustainable funds or portfolios might fail to deliver on their sustainable promises in the first place. One important reason is the economic incentives. Funds labeled as sustainable or ESG attract significantly higher capital inflows compared to similar conventional funds, even when the labeled funds’ aggregated sustainability characteristics are inferior compared to the conventional ones (Kaustia and Yu, 2021). This flow premium creates incentives for fund managers to adopt green labels without necessarily changing their investment process, which is a form of greenwashing, broadly defined as the practice of making unsubstantiated or misleading environmental claims (Lublóy et al., 2025; Delmas and Burbano, 2011). Additionally, Cao et al. (2026) find that funds that engage in greenwashing are able to charge higher fees while investors are more willing to accept weaker performance. This suggests that fund managers could exploit the positive association between sustainability labels and investor preferences without bearing the full costs of implementing sustainable investment practices, such as conducting detailed ESG research.

It is unknown how large the actual issue of greenwashing is. Dumitrescu et al. (2025) find that about 30% of ESG open-end mutual funds make explicit green claims that are not supported by fund activities. Popescu et al. (2023) document that 38% of SRI funds are more exposed to carbon-intensive firms than the MSCI World Index. D'Ecclesia et al. (2022) evaluate the environmental performance of ETFs and document poor performance. Similarly, Mazzolini et al. (2025) find that while the greenest funds invest more in low-carbon sectors, their carbon footprints remain comparable to conventional funds. Conversely, Meyers et al. (2024) report better environmental performance in terms of sustainability ratings[10] for sustainable funds.

Another challenge in identifying greenwashing lies in the two key metrics used for its measurement. These are ESG scores, with often a focus on the environmental pillar, and carbon emissions. As discussed in the previous section, ESG ratings diverge across providers (Berg et al., 2022). This divergence is not limited to the overall ESG scores but extends to the environmental pillars. Rating providers differ in which environmental dimensions they include, how they weight them, and how they measure firm-level performance on each dimension (MSCI, 2024b; LSEG, 2024a; FTSE Russell, 2025).[11] Carbon metrics, similarly, while seemingly more objective, can also face measurement challenges. Firms' emissions can be expressed in different ways, for example, as absolute emissions, emissions relative to revenue, market capitalization, or enterprise value including cash (EVIC). The final choice of a metric

[10] The term *sustainability ratings* (or *sustainability scores*) refers collectively to the combined ESG rating as well as its three pillar scores: Environmental (ENV), Social (SOC), and Governance (GOV). Much of the existing literature uses the label "ESG" to refer to all four rating types. To avoid implying that the analysis focuses only on the combined measure, Paper 4 adopts the umbrella term *sustainability ratings* to encompass any of these four scores.

[11] For example, MSCI structures its Environmental pillar into four themes (Climate Change, Natural Capital, Pollution & Waste, Environmental Opportunities), LSEG (formerly known as Refinitiv) uses three categories (Resource Use, Emissions, Innovation) and FTSE Russell uses five themes (Biodiversity, Climate Change, Pollution & Resources, Supply Chain: Environmental, Water Security). All three providers utilize industry-specific weighting, but differ in the industry classification system used and how they determine the final weights (MSCI, 2024b; LSEG, 2024a; FTSE Russell, 2025).

is important as it can affect the outcome of the sustainability evaluation (Gaskell, 2025; Shemfe et al., 2024). Furthermore, carbon data availability and quality are additional challenges. Carbon emissions consist of Scope 1 (direct emissions), Scope 2 (indirect emissions) and Scope 3 emissions (indirect emissions along the value chain).[12] While Scope 1 and 2, when available, are relatively accurate and have a high cross-provider consistency (Heldmann et al., 2025), Scope 3 emissions are often estimated and subject to greater measurement error and provider disagreement (Popescu et al., 2023; Busch et al., 2022). Moreover, emission data is not available for all firms, and the coverage varies across company size and time period (Heldmann et al., 2025). However, even for firms which report their data, some studies raise concerns over their reliability (Bajic et al., 2023). In order to achieve full emission coverage, data providers offer model-based estimates, using various methods, for companies which do not report their emissions (LSEG, 2024b; MSCI, 2024c).

The complexity of measuring both financial and sustainability outcomes thus presents an ongoing challenge for empirical research in this field. The following section introduces the research questions that guide this dissertation's contribution to this literature, along with the empirical strategies employed to address these challenges.

1.4 Overall Research Aim and Research Questions

The preceding discussion has established that two fundamental questions surround the sustainable finance literature, namely whether sustainable investment strategies affect financial performance and whether they achieve their non-financial objectives. However, the empirical evidence on both questions remains contested. This dissertation contributes to these debates through four empirical papers that examine different dimensions of sustainable equity investing. The overarching goal is to provide empirical evidence on both the financial and sustainability performance of sustainable investment strategies, while also addressing the methodological challenges that have contributed to the inconclusive findings in prior research.

Table 2 provides an overview of how the four papers map onto these two fundamental questions. Papers 1 and 2 address both questions by examining the same set of sustainable equity indices from financial and sustainability perspectives, respectively. Paper 3 investigates how much decarbonization is achievable through active portfolio optimization in European

[12] Direct Scope 1 emissions are from sources owned or controlled by the company. Indirect Scope 2 emissions are from purchased electricity, heat, or steam consumed at the facility. Indirect Scope 3 emissions cover upstream activities such as contractor vehicles, employee travel, and waste disposal, as well as downstream activities such as product use by customers and purchased materials production. These definitions follow Refinitiv Datastream variable descriptions (Scope 1: ENERDP024, Scope 2: ENERDP025, Scope 3: ENERDP096).

equities. Paper 4 tests whether sustainability ratings contain useful information for portfolio performance and factor models, while also examining the methodological sensitivity of sustainability-return relationships.

Table 2. Mapping of dissertation papers to research questions and strategies

Paper	Title	Question 1: Financial Performance	Question 2: Sustainability Effectiveness	Sustainable Investing Strategy
1	Financial Returns of Going Green: Evidence from MSCI Indices	✓		Exclusionary, Selective, Integrative
2	Investing in a Low-Carbon Transition: Carbon Footprint Savings with Green MSCI Indices		✓	
3	Small Deviations, Large Reductions: An Optimization Approach to Low-Tracking Error Decarbonized European Equity Portfolios	(✓)	✓	Integrative (Optimization)
4	Sustainability Ratings, Equity Portfolio Performance, and Factor Models: Evidence from a Multi-Specification Approach	✓		Selective (Best-in-Class)

Papers 1 and 2 examine the same set of 20 MSCI equity indices. These are four different sustainable index variants and one standard (parent) index in four different regions, namely the World, the USA, Europe, and Emerging Markets (EM).[13] The four sustainable MSCI index types represent varying strict[14] sustainability criteria implemented with multiple strategies: the Climate Change (CC) index, the Paris-Aligned Benchmark (PAB) index, the Socially Responsible Investment (SRI) index, and the SRI Filtered PAB (SRI PAB) index. This allows for a systematic comparison of how different index strategies and levels of sustainable strictness, which is called Green Criteria Effectiveness (GCE) in the paper, affect both financial returns and carbon emissions. Paper 1 addresses the financial performance questions by answering the following research questions:

> *1a. How do sustainable MSCI indices perform financially compared with their standard indices?*

> *1b. How does an index's Green Criteria Effectiveness (GCE) affect its financial performance?*

These questions are motivated by the observation that while numerous studies have compared sustainable and conventional indices, very few have directly examined how the construction

[13] In Papers 1 and 2, the sustainable index variants are referred to as *green* indices, used as an umbrella term. To maintain consistency throughout this dissertation, they are referred to as *sustainable*, except when directly referencing the Green Criteria Effectiveness (GCE) regression framework. Importantly, green in GCE can be understood as synonymous with sustainable in this context.

[14] In this context, strict describes how many stocks the index excludes compared with the parent index.

criteria of sustainable indices relate to their financial performance. If the criteria employed by MSCI do not affect a sustainable index's performance, any observed outperformance relative to its standard non-sustainable index is likely due to chance rather than the index construction methodology. Paper 2 addresses the sustainability effectiveness question under the same framework by answering the following two research questions:

2a. How do sustainable MSCI indices perform environmentally compared with their standard indices?

2b. How does an index's Green Criteria Effectiveness (GCE) affect its environmental performance?

These questions respond to concerns about greenwashing and the limited evidence on whether green investment products deliver measurable environmental improvements. By using carbon emissions data (Refinitiv, Bloomberg, and the Carbon Disclosure Project) and ESG scores (Refinitiv, Bloomberg, MSCI) from multiple providers, Paper 2 tests whether the sustainable label translates into genuine carbon reductions. Together, Papers 1 and 2 evaluate whether MSCI's sustainable index products deliver on both the financial and environmental dimensions.

Paper 3 shifts from evaluating existing sustainable index products to investigating what is technically achievable through active portfolio construction. Portfolio optimization allows investors to explicitly minimize carbon metrics subject to constraints on tracking error, sector weights, and other portfolio characteristics. This allows for the creation of a portfolio that assesses how much decarbonization is possible without sacrificing diversification or deviating substantially from its benchmark. The paper asks:

3. How much can portfolio-level carbon emissions be reduced in a European equity universe while maintaining a low tracking error relative to its parent index?

Most existing optimization-based decarbonization studies focus on global benchmarks, rely on estimated emissions data, and use proprietary commercial risk models that are difficult to replicate. Motivated by these gaps, Paper 3 focuses explicitly on European equities (the STOXX Europe 600), analyzes Scope 1–2 emissions and various carbon metrics, relies only on reported emissions from Refinitiv, and uses a transparent optimization framework.

Paper 4 returns to the financial performance question but takes a different approach. Rather than examining index products or optimized portfolios, it tests whether sustainability ratings from major providers contain useful information for portfolio performance and asset pricing models. This question is fundamental because sustainability ratings serve as the primary input for many sustainable investment strategies (Index Industry Association, 2023; Bell and Taylor, 2024). The paper addresses two related research questions:

4a. Do portfolios and factors with higher sustainability rating scores generate a higher (risk-adjusted) return than portfolios and factors with lower sustainability ratings?

4b. 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 these questions, Paper 4 employs a multi-specification approach that systematically varies portfolio construction decisions and examines how these choices affect financial performance and factor model outcomes. Using sustainability ratings from five major providers (Refinitiv, MSCI, RobecoSAM, Bloomberg, and Sustainalytics) covering four sustainability rating dimensions (Environmental, Social, Governance, and combined ESG) over 2003–2024, the paper provides a comprehensive assessment of whether sustainability ratings serve as useful investment signals. The paper also replicates and extends findings from three highly cited studies (Avramov et al., 2022; Pástor et al., 2022; Gibson-Brandon et al., 2021) to assess their robustness to alternative specifications and extended sample periods.

Each paper addresses the two fundamental questions from a different angle. Papers 1 and 2 evaluate what investors actually receive from commercially available sustainable index products. Paper 3 establishes what is technically achievable through optimization, providing a benchmark against which commercial products can be assessed. Paper 4 examines sustainability ratings as common building blocks of sustainable investment strategies and tests whether mixed findings in prior research reflect genuine economic relationships or methodological artifacts. Together, the papers cover multiple geographies (global regions in Papers 1–2, Europe in Paper 3, United States in Paper 4), investment approaches (passive index tracking, portfolio optimization, rating-based sorting), and outcome dimensions (financial returns, risk-adjusted performance, factor model enhancement, sustainability ratings, carbon emissions). This breadth allows the dissertation to offer a more complete picture of whether and how sustainable investment strategies can achieve both financial and sustainability objectives.

The remainder of this dissertation proceeds as follows. Chapter 2 provides a research overview presenting the data, methods, and key findings of each paper, with Papers 1 and 2 discussed jointly given their shared dataset, followed by Paper 3 and Paper 4. Chapter 3 concludes the dissertation by synthesizing findings across all four papers, discussing limitations and directions for future research, and offering concluding remarks. The four papers are included in their published or most recent working paper version in the Appendix.

2 Research Overview

2.1 Sustainable Indices: Financial and Sustainable Performance (Papers 1 & 2)

MSCI's sustainable equity indices represent one of the most important commercial implementations of sustainable investing principles.[15] Major asset managers like iShares, Amundi, or Xtrackers license these indices to construct Exchange-Traded Funds (ETF), which allow institutional but especially retail investors to cheaply invest broadly into a sustainable equity product. Papers 1 and 2 examine whether these indices deliver financially and sustainably, contributing to the debate outlined in Section 1.3. Additionally, by distinguishing between index variants with varying degrees of strictness, it can be tested how the heterogeneity of sustainable investing strategies documented in Section 1.2 influences the results.

While the broader ESG and SRI literature is extensive, four specific gaps within the sustainable index (or ETF) product field motivate these papers. First, prior research showed mixed evidence regarding financial performance, ranging from positive (Fiordelisi et al., 2023; Rompotis, 2023; Pavlova and de Boyrie, 2022), to mostly neutral (Bolognesi et al., 2024; Jonwall et al., 2024; Dumitrescu et al., 2023; Cunha et al., 2020; Jain et al., 2019), and, in some earlier studies, to negative results (Lean and Nguyen, 2014). The results regarding sustainability are similarly mixed, with some studies reporting poor or no improvement in environmental performance (D'Ecclesia et al., 2022; Abouarab et al., 2025), while others find that they generally outperform their conventional peers (Meyers et al., 2024; Jacob and Wilkens, 2021). Second, most studies treat different sustainable products as one category, which can obscure important differences between index construction methodologies which range from mild exclusion screens to strict Paris Agreement alignment. Third, the Climate Paris-Aligned Benchmark and SRI Filtered PAB indices, which were explicitly designed to meet Paris Agreement decarbonization targets, have not been empirically examined for their financial or environmental performance. Lastly, prior research predominantly examines ETFs rather than indices. Depending on the provider, ETFs have different expense ratios, replication strategies, or tax effects which can affect financial performance.

[15] Sustainable ETFs benchmarked to indices provided by MSCI represented about 75% of the market as of April 2024 (Heldmann et al., 2025).

Building on these gaps, the first paper addresses the financial dimension[16]:

> *1a. How do sustainable MSCI indices perform financially compared with their standard indices?*
>
> *1b. How does an index's Green Criteria Effectiveness (GCE) affect its financial performance?*

Paper 2 asks the parallel questions for sustainability outcomes:

> *2a. How do sustainable MSCI indices perform environmentally compared with their standard indices?*
>
> *2b. How does an index's Green Criteria Effectiveness (GCE) affect its environmental performance?*

These research questions are answered by analyzing 20 MSCI equity indices across four regions: World, U.S., Europe, and Emerging Markets. Within each region, the standard (parent) MSCI index serves as the benchmark against four sustainable variants: Climate Change (CC), Paris-Aligned Benchmark (PAB), Socially Responsible Investment (SRI), and SRI Filtered PAB (SRI PAB). These four variants differ in their strategies, ranging from mild exclusion criteria (CC excludes firms violating UN Global Compact principles and applies carbon-based reweighting) to the most restrictive (SRI PAB combines best-in-class ESG selection with Paris Agreement carbon reduction targets and a 5% issuer cap).[17] Compared to their parent index, the CC, PAB, SRI, and SRI PAB exclude on average 6%, 39%, 77%, and 79% of constituents, respectively.

MSCI provided monthly constituent weighting data from January 2015 to December 2022 for all 20 indices, comprising 5,267 constituents[18] across 50 countries and 11 sectors, resulting in 1,076,340 usable firm-month observations. Index-level return data extends through December 2023. Firm-specific financial data is obtained from Refinitiv Eikon. Carbon emissions data is sourced from three providers: Refinitiv (primary), Bloomberg, and CDP. ESG and environmental ratings are also obtained from three providers: Refinitiv (primary), Bloomberg, and MSCI. The use of three independent sources addresses reliability concerns noted in the literature regarding carbon and ESG data (Busch et al., 2022; Berg et al., 2022).

[16] The research questions in Papers 1 and 2 use the term *green* to describe the sustainable MSCI indices. For consistency with the terminology used throughout this dissertation, the term *sustainable indices* is used in the main text. The only exception is Green Criteria Effectiveness (GCE), which is retained from the papers, although *green* is interchangeable with *sustainable* in this context.
[17] For a detailed description of the construction methodology of each index variant, see Paper 1, Section 2.
[18] Of these, 5,034 constituents could be mapped to ISINs, representing 99.98% of total index weights.

Using this data, two primary outcome measures to assess both financial and sustainable performance are constructed. Financial performance is measured through the Cumulative Return Differential (CRD), defined as the accumulated monthly excess return of a sustainable index relative to its parent index over the investment horizon. Environmental performance uses size-adjusted Scope 1+2 emissions from Refinitiv, constructed for each index-month by weighting each constituent’s emissions scaled by market capitalization. For each primary measure, a battery of robustness tests is also carried out.

Clear patterns are observed for both dimensions with the primary performance measures. Figure 2 presents the average cumulative return differential (ACRD) across the four regions from January 2015 through December 2023. All four sustainable index variants outperform their standard counterparts over this period, with the SRI index achieving the highest ACRD of 13.7% (13% net of ETF costs), followed by SRI PAB at 7.5%, PAB at 5.8%, and CC at 4.7%. While all variants move rather closely until 2019, the stricter indices (SRI and SRI PAB) gain momentum in the latter years of the sample. Regional performance varies, with the SRI PAB delivering the strongest results in Europe (17.3% CRD), while the SRI performs best in Emerging Markets (15.6%). However, some variants slightly underperform in specific regions, most notably the CC, PAB, and SRI PAB in Emerging Markets, which have negative CRDs (−0.4% to −3%) over the full period.

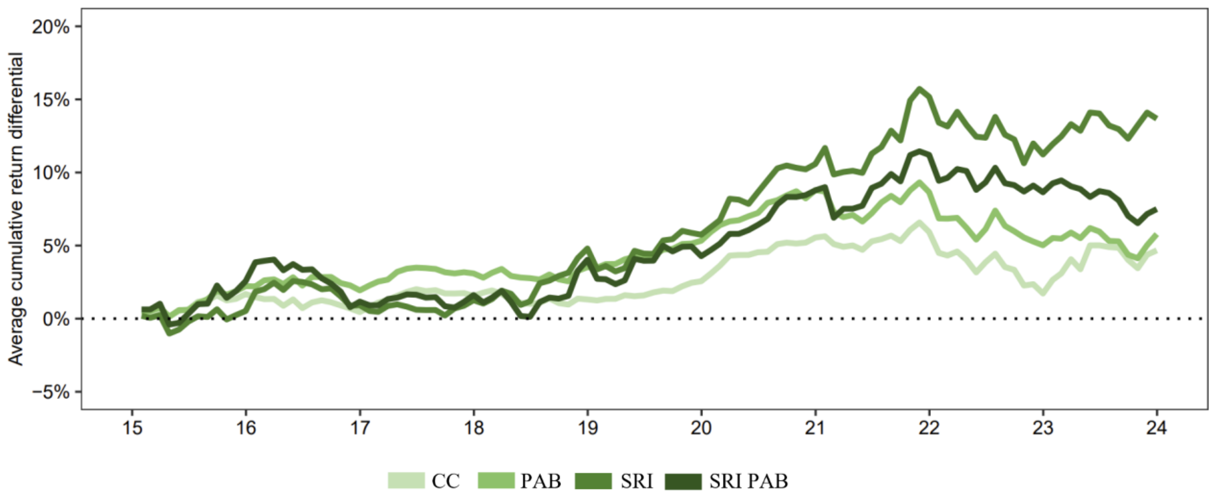


Figure 2. Financial performance of sustainable MSCI indices

Notes: This figure presents the average cumulative return differential (ACRD) of four MSCI sustainable index variants relative to the MSCI Standard index, averaged across four regions (World, USA, Europe, and Emerging Markets). The ACRD reflects the accumulated excess return of each sustainable index over its parent index from January 2015 through December 2023. For further details on the data and calculation methodology, see Paper 1, Section 4. CC = Climate Change; PAB = Paris-Aligned Benchmark; SRI = Socially Responsible Investment; SRI PAB = SRI Filtered PAB. Returns are gross of ETF expense ratios.
Source: Paper 1, Figure 1, Panel A.

These return patterns are supported by risk-adjusted measures such as Sharpe and Treynor ratios, as well as other financial metrics including Wealth Relative and Tobin's Q, all of which similarly favor the sustainable variants over their standard counterparts.[19] Moreover, time-varying correlations and market betas estimated from multiple Fama-French factor models remain close to one, suggesting that the sustainable indices are well diversified relative to their parent indices despite excluding up to nearly 90% of constituents in some cases.[20] This alleviates the concerns of potential suboptimal portfolio allocation and insufficient diversification implied by traditional portfolio theory (see Section 1.3.1).

Figure 3 shows the cumulative size-adjusted Scope 1+2 carbon emissions from Refinitiv for each region, assuming a US$1,000 investment at the start of the study period. Across all regions, it can be observed that the standard indices accumulate substantially higher emissions than their sustainable counterparts, with the gap widening consistently over time. Absolute emission levels differ considerably by region with the highest carbon footprint in Emerging Markets and the lowest in the USA. However, within each region, the relative ordering of index variants remains consistent.

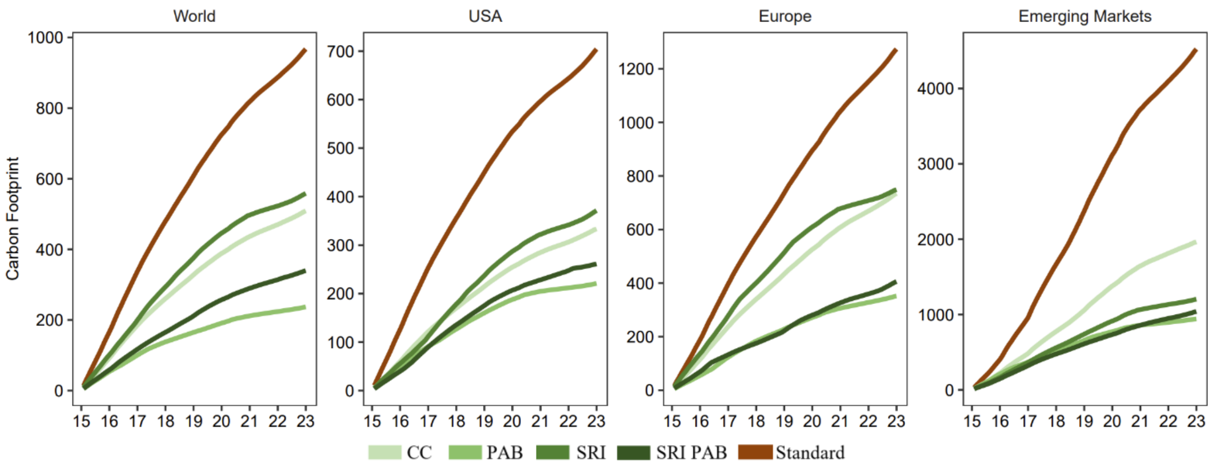


Figure 3. Carbon footprint over time by index type and region

Notes: This figure shows cumulative size-adjusted Scope 1+2 CO_2 emissions (in kg) for a US$1,000 investment in each index at the beginning of the study period (January 2015 to December 2022), using Refinitiv data. Emissions are scaled by market capitalization to yield emissions per dollar invested. For further details on the data and calculation methodology, see Paper 2, Section 3. The four regions shown are World, U.S., Europe, and Emerging Markets. CC = Climate Change; PAB = Paris-Aligned Benchmark; SRI = Socially Responsible Investment; SRI PAB = SRI Filtered PAB. Returns are gross of ETF expense ratios.
Source: Paper 2, Figure 3, Panel A, Row 1.

By December 2022, a US$1,000 investment in the standard index would have accumulated 1,867 kg of CO_2 on average across all regions, compared to 885 kg for CC (−53%), 720 kg for SRI (−61%), 512 kg for SRI PAB (−73%), and just 438 kg for PAB (−77%). The PAB and SRI

[19] See Paper 1, Table 2 and Paper 1, Table 3.
[20] See Paper 1, Table 4 and Paper 1, Figure 2.

PAB achieve the lowest emissions due to their explicit Paris Agreement alignment. These patterns are robust to alternative data sources and emission scope definitions. Bloomberg and CDP yield qualitatively similar results, as do analyses based on the individual Scope 1 and Scope 2 components rather than their sum.[21]

These descriptive results suggest that sustainable indices deliver on both financial and environmental dimensions, however, they do not control for compositional differences across indices. This question is addressed with a central methodological contribution, a regression framework that directly tests whether MSCI's index construction criteria cause measurable performance differences. In Papers 1 and 2, this is termed the Green Criteria Effectiveness (GCE), although the word green is interchangeable with sustainable in this context. Rather than simply comparing sustainable versus standard index performance (sustainable or financial), the GCE approach tries to isolate the incremental effect of each sustainable index type through dummy variables in an ordinary least squares (OLS) regression:

$$Performance_{i,t} = \alpha_0 + \beta_1 GCE_{CC} + \beta_2 GCE_{PAB} + \beta_3 GCE_{SRI} + \beta_4 GCE_{SRI_{PAB}} \qquad (1)$$
$$+\gamma Controls + \epsilon_{i,t}$$

where *Performance* can either be a financial or sustainable metric. The standard index serves as the baseline and is captured by the intercept, so each GCE coefficient represents the marginal effect of that green index type's construction criteria relative to the conventional index. Controls include the index's sector composition (percentage of firms in each sector), development composition (percentage of firms by Human Development Index (HDI) category), market capitalization, crisis period dummies for COVID-19 and the Ukraine war, and a post-launch dummy capturing potential performance differences after an index's official introduction. All regressions use Newey-West standard errors to address autocorrelation and heteroskedasticity concerns.

Table 3 presents the baseline GCE regression results for both primary performance measures. On the financial side, all four GCE coefficients are positive and statistically significant at the 1% level, confirming that the sustainable index construction criteria generate higher cumulative returns relative to the standard index after controlling for compositional differences. The coefficients increase with the strictness of the green criteria, ranging from 2.0 percentage points for CC to 7.7 percentage points for SRI PAB. On the environmental side, all four GCE coefficients are negative and highly significant, indicating that the sustainable indices

[21] See Paper 2, Table 1 and Paper 2, Figure 3.

accumulate substantially lower carbon emissions per dollar invested. The PAB achieves the largest emission reduction (615.5 kg), followed by SRI PAB (587.6 kg), SRI (366.2 kg), and CC (303.9 kg). These results suggest that MSCI's sustainable index construction criteria are effective in delivering both superior financial returns and meaningful carbon emission reductions, even after accounting for various confounding factors. These baseline findings are robust across a battery of tests reported in Papers 1 and 2, including alternative performance measures, data providers, subsamples, sub-periods, and lag structures.[22]

Table 3. Financial and sustainable performance of sustainable indices

	Dependent Variable	
Independent Variable	Financial Performance: Cumulative Return Differential	Sustainable Performance: Refinitiv cum. size-adj. Scope 1+2
GCE: CC	0.02***	−303.9***
	(0.007)	(36.1)
GCE: PAB	0.053***	−615.5***
	(0.009)	(56.6)
GCE: SRI	0.059***	−366.2***
	(0.004)	(33.3)
GCE: SRI PAB	0.077***	−587.6***
	(0.009)	(50.6)
Sector composite	YES	YES
HDI composite	YES	YES
Log of market value	YES	YES
Post Launch dummy	YES	YES
Ukraine dummy	YES	YES
Covid dummy	YES	YES
Observations	1,900	1,900
Adj. R2	0.528	0.740

Notes: This table reports the GCE coefficient estimates from OLS regressions of the cumulative return differential and cumulative size-adjusted Scope 1+2 CO_2 emissions in kg per US$1,000 invested on index-type dummy variables and controls over January 2015 to December 2022. The standard (parent) index serves as the baseline and is captured by the intercept. Emissions data is sourced from Refinitiv. Newey-West standard errors are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.
Source: The financial performance column is taken from Paper 1, Table 5, Col. 1. The environmental performance column is taken from Paper 2, Table 2, Col. 1.

While carbon emissions show the most direct environmental outcome, they represent only one part of sustainability. Environmental and ESG ratings, on the other hand, offer a broader assessment of sustainability. The SRI and SRI PAB indices consistently achieve the highest environmental and ESG ratings across all three providers (Refinitiv, Bloomberg, and MSCI). The CC and PAB indices, which do not have an explicit SRI focus, score closer to, or, depending on the provider, even below the standard index.

[22] For the full set of robustness tests, see Paper 1 (financial performance), Table 5 and Appendix B, and Paper 2 (environmental performance), Table 2, Appendix D, Appendix E, and Appendix F.

Table 4 analyzes these patterns with the regression framework using the main source Refinitiv. The SRI and SRI PAB show large and significant positive coefficients for both measures, which indicates that their best-in-class approach is able to identify firms with better sustainability characteristics.[23] The CC index also shows positive and significant effects, although at a lower magnitude. Only the PAB exhibits significant negative coefficients for both rating measures, and Bloomberg confirms qualitatively similar results across all index types.

Table 4. Regression results for environmental and ESG rating scores

Independent Variable	Dependent Variable: Sustainable Performance	
	Refinitiv Environmental Score	Refinitiv ESG Score
GCE: CC	1.65***	1.55***
	(0.27)	(0.3)
GCE: PAB	−2.22***	−1.89***
	(0.48)	(0.54)
GCE: SRI	4.44***	5.03***
	(0.2)	(0.21)
GCE: SRI PAB	3.78***	4.74***
	(0.31)	(0.39)
Sector composite	YES	YES
HDI composite	YES	YES
Log of market value	YES	YES
Post Launch dummy	YES	YES
Ukraine dummy	YES	YES
Covid dummy	YES	YES
Observations	1900	1900
Adj. R2	0.915	0.891

Notes: This table reports the GCE coefficient estimates from OLS regressions of Refinitiv's environmental rating score and Refinitiv's ESG rating score on index-type dummy variables and controls over January 2015 to December 2022. Rating scores range from 0 to 100, with higher values indicating better sustainability performance. The standard (parent) index serves as the baseline and is captured by the intercept. Newey-West standard errors are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.
Source: Paper 2, Table 3, Cols. 2 and 6.

To investigate this puzzling finding for PAB, two additional analyses are conducted in Paper 2. A cumulative weighting difference (CWD) analysis shows that, when ranked by Refinitiv or Bloomberg environmental ratings, the PAB index tends to overweight constituents with lower ratings and underweight those with higher ratings relative to the standard index.[24] An attribution analysis further decomposes the environmental rating into sector allocation and stock selection components and finds that both components contribute negatively for the PAB. However, under MSCI's own environmental ratings, all four sustainable index variants have positive excess

[23] This pattern is further confirmed by the GCE regression for Refinitiv's ESG Controversies score, where the SRI and SRI PAB show the largest and most significant positive coefficients, indicating fewer ESG-related controversies among their constituents (see Paper 2, Table 3, Col. 1).
[24] For calculation details and results, see Paper 2, Appendix G.

ratings driven by better sector allocation and stock selection.[25] This suggests that the negative PAB coefficients are partly attributable to divergences in ESG methodologies by the providers.

Papers 1 and 2 together provide an evaluation of sustainable indices with regard to financial and sustainable performance. On the financial side, the sustainable MSCI indices outperform their standard counterparts on average across all four variants over the sample period, although not across all regions (Question 1a). The GCE regressions confirm that this outperformance is statistically significant and linked to the index construction criteria, with all four GCE coefficients positive and significant after controlling for compositional differences (Question 1b). The concern implied by traditional portfolio theory that restricting the investable universe may negatively affect financial returns or diversification is not supported by the results. On the sustainability side, all sustainable index variants achieve lower carbon emissions than their standard counterparts, with the PAB and SRI PAB delivering the largest reductions due to their explicit Paris Agreement alignment (Question 2a). The GCE coefficients confirm that these reductions are statistically significant and linked to the index construction criteria (Question 2b). Beyond carbon emissions, the broader sustainability assessment through environmental and ESG ratings reveals a more differentiated picture. The SRI and SRI PAB consistently achieve the highest rating improvements across providers, while the PAB shows limited or even negative effects depending on the rating source.

The findings also show that the different labels and strategies characterized in Section 1.2 translate into meaningfully different outcomes. The least strict variant, the Climate Change Index, has modest financial outperformance and carbon reductions but delivers a limited improvement in broader sustainability ratings. The PAB achieves the strongest carbon emission reductions, consistent with its explicit Paris Agreement alignment, but shows negative effects on environmental and ESG ratings under Refinitiv and Bloomberg, although not under MSCI's own ratings. The SRI, which applies the best-in-class ESG selection, has the highest average financial outperformance and high rating improvements, but its carbon reductions, while substantial, are smaller than those of the Paris-aligned variants. The SRI PAB, despite being the most restrictive variant, appears to be the most balanced profile regarding high financial returns, large carbon emission reductions, and high improvements in broader sustainability ratings. Overall, the results suggest that sustainable index products offer both retail and institutional investors viable options for aligning their equity portfolios with sustainability objectives

[25] For calculation details and results, see Paper 2, Appendix H.

without systematically sacrificing financial performance, though the choice of variant involves trade-offs that depend on the investor's priorities.

These conclusions are subject to some limitations. The analysis covers 20 indices across four regions, which, while broad in scope, represents only a part of the index universe. Furthermore, the sample period of 2015 to 2022 (2023 for returns) is relatively short. Longer time horizons are needed to assess whether the observed patterns will hold.

2.2 Active Portfolio Construction for Decarbonization (Paper 3)

Papers 1 and 2 evaluate the financial and environmental performance of commercially available sustainable indices. Paper 3 asks a follow-up question, that is, how much investors can reduce carbon emissions by actively optimizing a portfolio for decarbonization while keeping the portfolio close to its benchmark.

This question is addressed by constructing low-carbon variants of the STOXX Europe 600, over the period January 2011 to December 2024. The existing portfolio optimization decarbonization literature has several gaps that motivate this study. First, most optimization-based studies focus on global benchmarks (Bouchey et al., 2024; Bender et al., 2019; Blitz et al., 2024).[26] Second, prior studies rarely separate Scope 1–2 from Scope 3 emissions in their optimization, even though Scope 3 data is typically estimated rather than reported and subject to much greater measurement error and provider disagreement (Popescu et al., 2023; Busch et al., 2022). Third, the construction of low-carbon portfolios in the literature almost always relies on carbon datasets that include estimated emissions, such as S&P Global Trucost (Jondeau et al., 2025; Bouchey et al., 2024; Fahlenbrach and Jondeau, 2023; Bender et al., 2019; Andersson et al., 2016). Instead, only reported Scope 1–2 emissions from Refinitiv are used. Fourth, many studies use proprietary commercial risk models and optimizers such as MSCI Barra (Bouchey et al., 2024; Andersson et al., 2016) or Axioma (Bender et al., 2020; Bender et al., 2019) to estimate ex ante tracking error and construct portfolios, whereas a self-built multi-factor risk model and optimizer is constructed.[27] With this approach, the paper aims to answer the following question:

> *3. How much can portfolio-level carbon emissions be reduced in a European equity universe while maintaining a low tracking error relative to its parent index?*

[26] Europe only appears as one region within a global portfolio (Jondeau et al., 2025) or in early work on MSCI Europe (Andersson et al., 2016).
[27] For a comparative overview of carbon reductions and tracking errors reported in these studies, see Paper 3, Online Appendix A.

The main carbon measure is Scope 1+2 carbon intensity, defined as the weighted average of each constituent's carbon emissions per unit of revenue.[28] The optimization minimizes this measure subject to an ex ante tracking error (TE) budget and a set of practical constraints on sector, country, and factor exposures, position sizes, and liquidity.[29]

Three different scenarios are constructed, which differ in how much deviation from the benchmark is allowed and get progressively looser. The Conservative portfolio uses an ex ante TE budget of 0.5% with tight sector, country, and factor exposure limits (±1%), the Balanced portfolio allows a 1.0% TE budget with moderate constraints (±2%), and the Aggressive portfolio has a TE up to 2.0% with wider bounds (±3%). These TE levels were selected from the carbon–tracking error frontier shown in Figure 4, which plots the average carbon-intensity reduction achievable at each level of realized tracking error.[30]

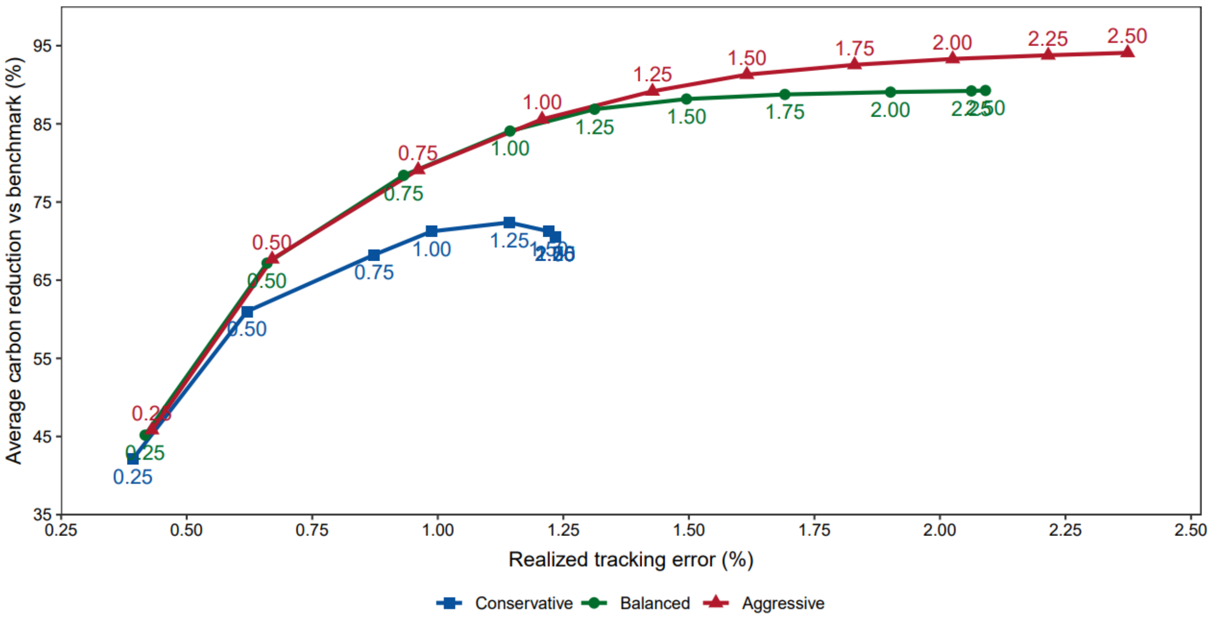


Figure 4. Carbon–tracking error frontier

Notes: This figure plots the average percentage reduction in portfolio Scope 1–2 carbon intensity relative to the STOXX Europe 600 benchmark against realized tracking error for three constraint sets (Conservative, Balanced, Aggressive). For each scenario, portfolios are optimized to minimize revenue-based Scope 1–2 carbon intensity under a grid of ex ante tracking-error budgets from 0.25% to 2.50%, holding all other constraints fixed. Realized tracking error and average carbon reductions are computed from monthly returns over 2011–2024.
Source: Paper 3, Figure 2.

The frontier shows a strongly concave relationship, carbon reductions increase steeply at low tracking error levels, but the marginal gain diminishes beyond approximately 2%. This means

[28] Paper 3 uses the notation "Scope 1–2" to denote the same measure. For consistency within the main text, "Scope 1+2" is used throughout the Research Overview. See Paper 3, Section 2.2.1 for the formal definition.
[29] See Paper 3, Sections 3.1–3.2 for the optimization framework and constraint set, and Online Appendix B for the risk model.
[30] See Paper 3, Table 2.

that a large share of the achievable decarbonization can be captured with only modest departures from the benchmark.

Table 5 reports the main carbon, risk, and performance outcomes. The key finding is that substantial Scope 1+2 reductions are achievable even at very low tracking error (Question 3). The Conservative portfolio reduces Scope 1+2 carbon intensity by 61% and carbon footprint by 58%, at realized TE of only 0.64%. The Balanced portfolio achieves an 84% intensity reduction and 79% footprint reduction for a realized TE of 1.17%. The Aggressive portfolio pushes intensity reductions above 93% with a higher realized TE at 2.08%. Importantly, although the optimization targets only Scope 1+2 emissions, all three portfolios also deliver substantial reductions in Scope 1+2+3 metrics (26% to 66% for carbon intensity), demonstrating co-benefits for value-chain emissions even when they are not directly optimized.

Table 5. Carbon, risk, and performance outcomes of optimized low-carbon portfolios

Metric	Benchmark	Conservative	Balanced	Aggressive
ex ante Tracking Error budget (%)	-	0.50	1.00	2.00
ex post Tracking Error realized (%)	-	0.64	1.17	2.08
Carbon Intensity Reduction (S1+2, %)	**0.0**	**61.0**	**84.1**	**93.3**
Carbon Footprint Reduction (S1+2, %)	0.0	57.5	79.3	93.3
Carbon Intensity Reduction (S1+2+3, %)	0.0	26.1	48.2	66.3
Carbon Footprint Reduction (S1+2+3, %)	0.0	20.8	40.3	64.5
Cumulative return (%)	196.1	209.2	230.8	267.5
Annual return (%)	8.06	8.40	8.92	9.74
Annual volatility (%)	13.31	13.31	13.26	13.37
Sharpe Ratio	0.61	0.63	0.67	0.73
Information Ratio	-	0.52	0.73	0.81

Notes: This table reports carbon reductions, tracking quality, and return metrics for three optimized low-carbon portfolios and the STOXX Europe 600 benchmark for the period 2011–2024. The optimization objective is to minimize revenue-based Scope 1+2 carbon intensity. Carbon reductions are expressed as percentage reductions relative to the benchmark. Returns and risk statistics are based on monthly total returns in euros, gross of management fees and transaction costs.
Source: Paper 3, Table 3.

There is no evidence of a performance penalty associated with decarbonization. Annualized returns increase slightly from 8.1% for the benchmark to between 8.4% and 9.7% across the three scenarios. Total volatility remains nearly identical at around 13.3%. Sharpe ratios rise from 0.61 to between 0.63 and 0.73, and information ratios are positive throughout, ranging from 0.52 (Conservative) to 0.81 (Aggressive). These results support the findings from Papers 1 and 2, that sustainable strategies do not need to sacrifice financial performance, although here it is demonstrated through active portfolio optimization rather than through existing index products. However, part of the observed outperformance may reflect favorable factor dynamics during the sample period, with the Balanced and Aggressive portfolios slightly tilting toward momentum and away from value, and both of these tilts were rewarded during 2011–2024. Additionally, a self-constructed low-carbon factor also showed positive returns during this period. Therefore, it remains an open question whether the positive information ratios would persist under different market regimes.

The baseline results are robust across a comprehensive battery of tests. Replacing Scope 1+2 carbon intensity with Scope 1+2 carbon footprint as the optimization objective, using EVIC instead of market value in the footprint denominator, extending the objective to Scope 1+2+3 metrics, switching to the broader STOXX Europe Total Market Index, winsorizing extreme carbon observations, restricting the sample to the post-Paris period (2016–2024) or to 2020–2024, and converting all returns to US dollars leave the main conclusions qualitatively unchanged. Across these specifications, large emissions reductions remain achievable at low tracking error, and risk-return profiles remain comparable to or slightly better than the benchmark. Additionally, net-of-cost information ratios remain positive for the Balanced and Aggressive portfolios (0.45 and 0.65, respectively), broader sustainability scores from Refinitiv and MSCI are maintained or modestly improved, and Fama-French factor exposures remain small. Only the Balanced and Aggressive portfolios show statistically significant tilts away from value and, in the case of the Aggressive portfolio, toward momentum.

Beyond aggregate carbon and financial performance outcomes, a question of practical relevance for investors is where these reductions actually come from. Specifically, whether they are driven by the allowed sector tilts or by within-sector stock selection, which essentially amounts to a best-in-class approach within each sector performed by the optimizer. Paper 3 answers this question by employing the Brinson attribution framework (Brinson et al., 1986), whose results are shown in Table 6. For carbon reductions (Panel A), within-sector stock selection accounts for the dominant share of the total reduction: 85% in the Conservative portfolio, 77% in the Balanced, and 67% in the Aggressive. Even in the most active portfolio, where sector tilts contribute about one-third of the emission savings, the optimizer primarily achieves decarbonization by replacing high-emitting stocks with lower-emitting ones within the same sectors. The performance decomposition in Panel B has a similar pattern. Within-sector selection explains 110%, 92%, and 74% of the cumulative active return in the Conservative, Balanced, and Aggressive portfolios, respectively, with sector allocation contributing only modestly.

Table 6. Decomposition of carbon reductions and active returns

Panel A: Contribution to Scope 1+2 Carbon-Intensity Reduction (% of Absolute Reduction)

Attribution	Conservative	Balanced	Aggressive
Sector tilts	15.4	22.6	33.0
Within-sector selection	84.6	77.4	67.0

Panel B: Brinson Attribution of Cumulative Active Return

Attribution	Conservative (% of active)	Balanced (% of active)	Aggressive (% of active)
Allocation	−7.9	10.4	21.1
Selection	109.9	91.9	74.0
Interaction	−2.1	−2.4	4.9

Notes: Panel A reports the mean share of sector tilts and within-sector stock selection in the absolute reduction of Scope 1+2 carbon intensity across portfolio years over the period 2011–2024. Panel B decomposes cumulative active return into allocation, selection, and interaction effects. Values in Panel B are expressed as percentages of total cumulative active return.
Source: Paper 3, Table 5.

These results have direct practical implications for asset managers seeking to decarbonize their (European) portfolios. The dominance of within-sector stock selection in both the carbon and return decompositions implies that managers can achieve the majority of emission reductions and active returns by maintaining sector exposures close to the benchmark and tilting toward cleaner companies within each sector. This reduces concerns about unintended style or sector bets that could arise from more aggressive reweighting strategies. The study is limited by its sample period of 2011–2024, which also coincides with generally favorable conditions for low-carbon strategies. Also, the reliance on annual rebalancing means that the results may not fully capture the costs and turnover associated with more frequent portfolio adjustments. Additionally, the framework is backward-looking in the sense that it optimizes on reported historical emissions rather than on forward-looking transition metrics or climate scenario alignment, which are increasingly relevant for investors with net-zero commitments (Cenedese et al., 2023; Bolton et al., 2022).

2.3 Sustainability Ratings as Investment and Pricing Signals (Paper 4)

Papers 1 and 2 analyze commercially available indices, for some of which sustainability ratings are a direct input in the construction process, determining index eligibility and constituent weightings. In Paper 3, portfolios are optimized solely on carbon emissions data, which, although only representing one dimension of the environmental pillar, resulted in low-carbon portfolios with similar or improved sustainability ratings as a byproduct of selecting cleaner firms within each sector. Paper 4 now turns to sustainability ratings directly as the object of analysis, asking whether these scores carry useful information for equity portfolio performance and factor models.

Sustainability ratings have become a central input in institutional investment practice (Bell and Taylor, 2024; Index Industry Association, 2023). Despite this widespread adoption, academic literature has not reached a definitive answer on whether sustainability ratings can be used to generate superior financial performance. Empirical studies on risk-adjusted returns of rating-sorted portfolios and factors report mixed findings. For example, Díaz et al. (2021), Madhavan et al. (2021), and Khan (2019) find evidence supporting positive risk-adjusted performance, while Ciciretti et al. (2023) document negative outcomes and Hübel and Scholz (2020) and Alves et al. (2025) report largely neutral results or little evidence of systematic outperformance. Halbritter and Dorfleitner (2015) and Dobrick et al. (2025) similarly observe neutral performance for ESG-sorted portfolios. The literature on whether sustainability-based factors can improve the explanatory power of traditional asset pricing models is equally divided, with Dobrick et al. (2025), Díaz et al. (2021), and Jin (2018) documenting improvements to Fama-French factor models, while Naffa and Fain (2022) and an older study by Xiao et al. (2013) report no significant enhancement.

The mixed findings in this literature are due, at least in part, to two limitations. First, a large share of studies[31] relies on a single rating provider, despite the well-documented evidence that sustainability ratings diverge substantially across providers (Berg et al., 2022). Second, and especially critical through its interaction with the first limitation, the methodological choices in portfolio construction introduce additional uncertainty. When researchers sort stocks into portfolios based on a sustainability characteristic, they face numerous decision points, including firm-size thresholds, stock-price filters, sector exclusions, weighting schemes, and sorting procedures, each of which can meaningfully affect the resulting performance estimates. Walter et al. (2024) refer to this problem as "methodological uncertainty" in the context of standard (non-sustainable) portfolio sorts, showing that seemingly small specification choices can produce widely divergent outcomes. In the context of sustainability ratings, where the signal itself is noisy and provider-dependent, this methodological uncertainty compounds the already mixed empirical picture. To date, only one other study has begun to quantify methodological uncertainty in sustainability rating-related portfolio construction.[32] This gap is addressed more comprehensively by analyzing ratings from five major providers (Refinitiv, MSCI,

[31] More recent studies have begun to address this limitation by employing multiple providers (e.g., Alves et al., 2025; Gibson-Brandon et al., 2021; Avramov et al., 2022).

[32] Henriquez-Salman (2025) also quantifies methodological uncertainty in ESG portfolio sorts. However, this study differs in its more extensive data coverage and methodological design, incorporating additional decision nodes such as stock-price filters, earnings and book-value thresholds, and exchange selection for breakpoints. Importantly, the rebalancing frequency is also adapted to reflect sustainability data publication cycles, using annual rebalancing with appropriate lags rather than the monthly rebalancing employed in Henriquez-Salman (2025).

RobecoSAM, Bloomberg, Sustainalytics) across 4,481 U.S. firms over 21 years (2003–2024). In addition, two cross-provider measures are constructed, which are rating Agreement (the mean of normalized scores across providers) and rating Disagreement (their standard deviation).[33] Together, they capture the consensus and divergence among providers for each firm. Beyond this broad data coverage, specification choices are systematically varied across multiple decision nodes to assess the sensitivity of results to portfolio construction decisions. The analysis further includes replications of three influential studies, factor spanning tests, maximum ex-post Sharpe ratios, and GRS tests (Gibbons, Ross, and Shanken, 1989). With this broad data coverage and multi-specification design, Paper 4 addresses two questions:

> *4a. Do portfolios and factors with higher sustainability rating scores generate a higher (risk-adjusted) return than portfolios and factors with lower sustainability ratings?*
>
> *4b. 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?*

The methodological foundation of the paper is a multi-specification framework for portfolio construction, inspired by the approach of Walter et al. (2024) for standard anomalies. The adopted framework systematically varies portfolio building decisions at each of the following nodes: (1) whether to exclude firms with market capitalization below US$300 million, (2) whether to exclude stocks priced below US$5, (3) whether to exclude firms with negative book value, (4) whether to exclude firms with negative earnings, (5) whether to exclude financial firms, (6) whether to exclude utility firms, (7) the lag structure between rating observation and portfolio formation, (8-12) double-sorting parameters including sort dependence, breakpoint determination, and sorting granularity, and (13) value-weighting versus equal-weighting.[34] This procedure generates 128 distinct specifications for one-way-sorted decile portfolios and 512 specifications for two-way-sorted portfolios based on ratings and firm size. Rating data, updated annually, covers the period 2002 to 2022, with portfolio formation following the standard Fama-French (1993) timing convention, that is, ratings observed in December of year t-1 are used to form portfolios in June of year t for the period July t to June t+1.

Before the original empirical analysis, the relevant portfolio and factor results from three highly cited papers are replicated, each examining a different aspect of the sustainability rating-return relationship. The replications employ both the single-specification method documented in each original study and the multi-specification framework.

---

[33] Both measures require ratings from at least three of the five providers. See Paper 4, Section 4.1 for the formal definitions.

[34] See Paper 4, Figure 1.

The first replication is of Avramov et al. (2022), who report that when ESG rating uncertainty among providers is low, brown (low-rated) stocks outperform green (high-rated) stocks. Using 512 specifications, the replication finds that the long-short portfolio returns and alphas reported by Avramov et al. (2022) are mostly negative and rarely statistically significant, with the median multi-specification estimates being consistently smaller and mostly negative compared to the original study's positive point estimates. The range of alpha estimates is wide, spanning from negative to positive values, indicating strong sensitivity to portfolio construction choices.[35]

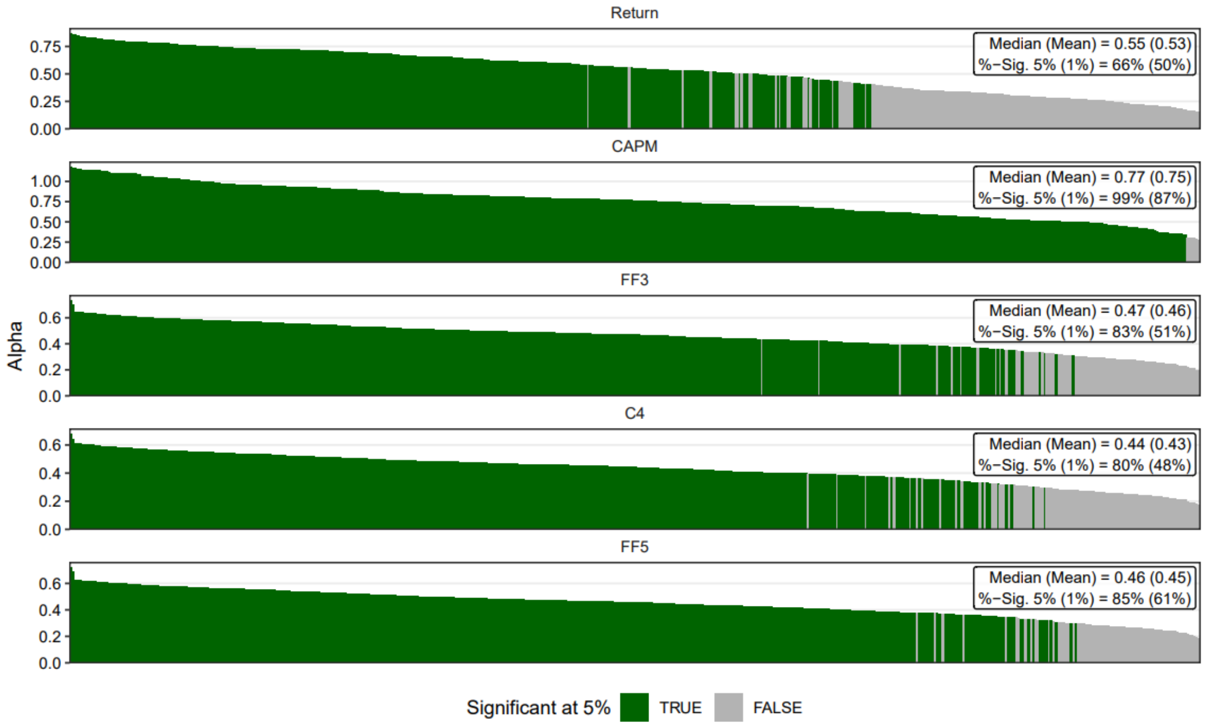
The second replication addresses Pástor et al.'s (2022) Green-Minus-Brown (GMB) factor, which documents significant green outperformance during November 2012 to December 2020 for MSCI's environmental score.[36] Figure 5 illustrates how the multi-specification analysis affects these findings. During the original time window, the majority of alpha estimates remain statistically significant across factor models, with between 66% and 99% of specifications producing significant alphas at the 5% level (Panel A). Extending the sample to June 2024, however, causes the median alphas to decline by roughly two-thirds and the proportion of significant specifications to collapse to between 0% and 27% depending on the factor model (Panel B).

The third replication concerns Gibson-Brandon et al. (2021), who document positive return premiums for portfolios sorted on industry-adjusted rating disagreement among S&P 500 stocks during 2010–2017. The multi-specification analysis reveals that the majority of alpha estimates are statistically insignificant in both the original and extended periods, with fewer than 10% of specifications producing significant alphas for ESG and ENV in the extended period 2010–2023.[37]

The overall observation across the three replicated studies is that results appear robust under a single specification. However, these results become fragile and less clear after analyzing the full range of plausible portfolio construction options and extending their sample periods.

[35] See Paper 4, Figure 2 for the full distribution of multi-specification alpha estimates. Details on the replication procedure are in Paper 4, Online Appendix A.
[36] For details on the replication procedure, see Paper 4, Online Appendix B.
[37] The multi-specification results are in Paper 4, Figure 4, and the single-specification replication is in Online Appendix Table C.5. Replication details are in Online Appendix C.

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



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

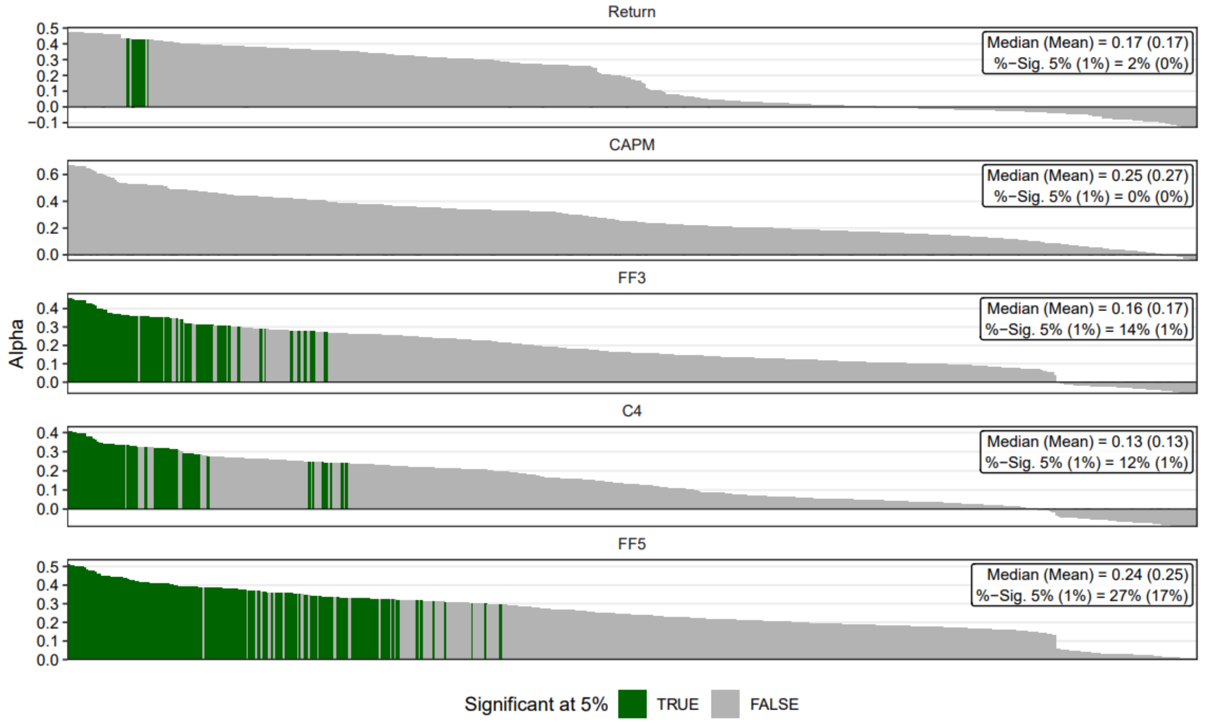


Figure 5. Pástor et al. (2022) replication: alpha estimate distribution

Notes: This figure shows the distribution of 512 alpha estimates corresponding to Pástor et al. (2022) Table 3 for five factor models: Return (model with constant), CAPM, Fama-French three-factor (FF3), Carhart four-factor (C4), and Fama-French five-factor (FF5). The dependent variable is the respective GMB spread specification. Bars are sorted from highest to lowest alpha, with green indicating significance at the 5% level and grey indicating insignificance. Median and mean alphas and proportions of significant alphas (at 5% and 1%) are reported in each plot. All models are estimated with Newey-West standard errors. Panel A covers the original study period ending December 2020; Panel B extends to June 2024.

Source: Paper 4, Figure 3.

After having established that existing results are sensitive to specification choices, the analysis turns to the full sample to answer the question whether there is any robust rating-return relationship when applying the multi-specification framework.

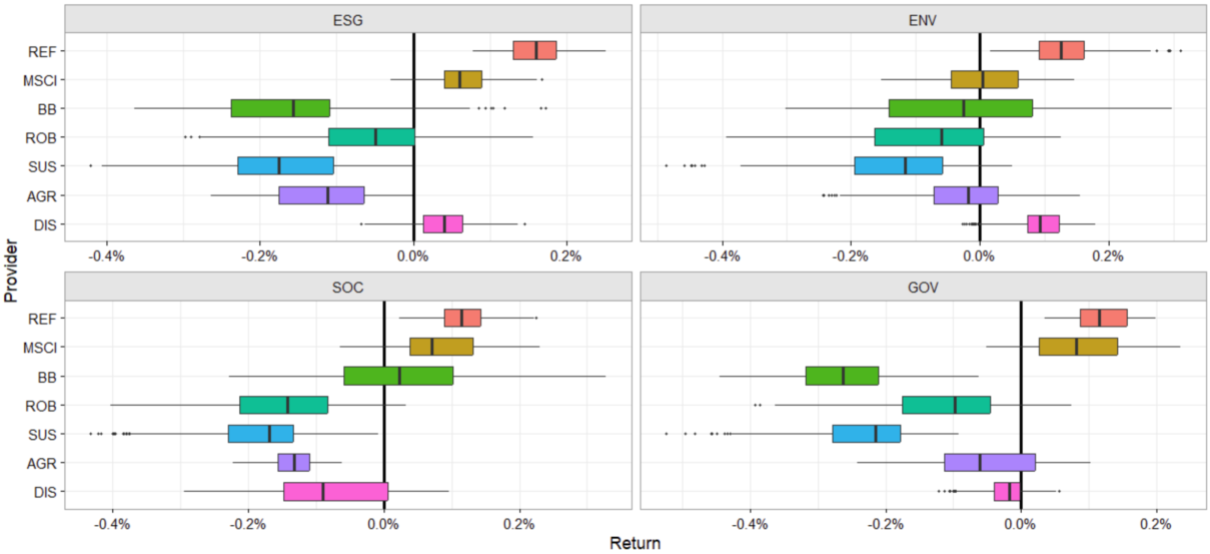
First, one-way-sorted decile portfolios are analyzed. For these, no clear and consistent pattern linking higher sustainability rating scores to better financial performance is observed. Median excess returns fluctuate across portfolios without exhibiting a monotonic upward trend, and this finding holds across providers and rating categories. Adjusting for volatility through Sharpe ratios results in somewhat more favorable results for higher-rated portfolios in some cases, but recognizable patterns are still not observed consistently across providers and rating categories.[38]

Since firms with larger market capitalization typically exhibit higher ESG scores[39], a concern is that the observed return patterns may be confounded by a size effect. To disentangle the two, double-sorted portfolios based on ratings and firm size are constructed. These double sorts similarly fail to establish a consistent positive relationship between ratings and financial performance. While larger firms tend to carry higher ratings, this does not translate into consistently better returns for highly rated portfolios after controlling for size.[40]

This relationship is further examined by constructing rating-based long-short factors as the average return difference between the two high-rated portfolios (Big-High and Small-High) and the two low-rated portfolios (Big-Low and Small-Low) for each provider and rating type, generating 512 factor specifications per provider-rating combination. Figure 6 shows the distribution of returns (Panel A) and Sharpe ratios (Panel B) for each rating factor across providers.

[38] See Paper 4, Appendix B for the full set of excess return (Panel A) and Sharpe ratio (Panel B) boxplots across providers and rating categories.
[39] Across all five providers, median normalized ESG scores are substantially higher for firms in the top market-capitalization quartile than in the bottom, with differences ranging from +10 (MSCI) to +31 (RobecoSAM) points. See Paper 4, Online Appendix Table D.1 for summary statistics of normalized rating scores across providers.
[40] See Paper 4, Appendix C for heatmaps of median excess returns (Panel A) and Sharpe ratios (Panel B) of the 4×4 rating×size portfolios.

Panel A: Factor Returns



Panel B: Sharpe Ratios

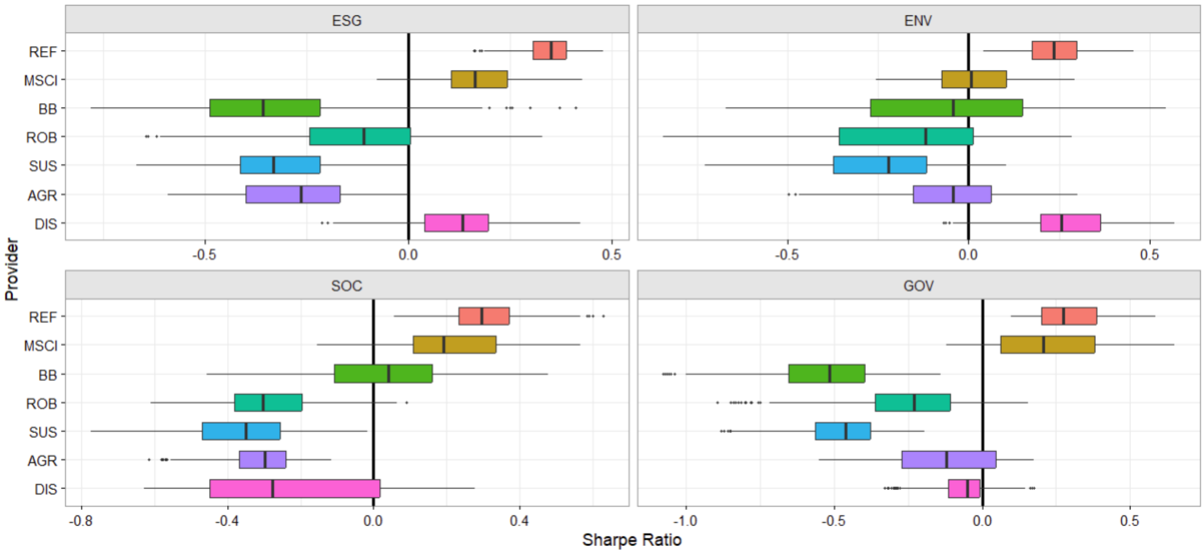


Figure 6. Returns and Sharpe ratios of rating factors

Notes: This figure depicts the distribution of rating factor returns (Panel A) and Sharpe ratios (Panel B) across 512 specifications for each provider and rating type. The providers are 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 the self-calculated rating Agreement (AGR; July 2010–June 2024) and Disagreement (DIS; July 2010–June 2024). Each factor is constructed as the average return difference between the two high-rated portfolios (Big-High and Small-High) and the two low-rated portfolios (Big-Low and Small-Low). Boxplots show the median, interquartile range, and the full range of specification outcomes. A box entirely in positive territory indicates that at least 75% of specifications produce positive values.
Source: Paper 4, Figure 5.

Most factors generate neutral to negative median returns and Sharpe ratios, with only Refinitiv and, to a lesser extent, MSCI producing consistently positive values. However, even for these two providers, statistical significance is rare, with the share of significant specifications ranging from 1% to 19% for Refinitiv depending on the rating dimension. Furthermore, individual specification choices can substantially shift factor returns. For instance, choosing value-weighting over equal-weighting alone changes median monthly Bloomberg ENV factor returns by 0.26 percentage points, and excluding firms with negative earnings shifts Bloomberg GOV factor returns by 0.21 percentage points.[41]

Similarly, when examining cumulative factor returns over the full sample horizon, the variation within a single provider-rating combination can be substantial. For example, the Refinitiv ENV factor's cumulative returns range from 1.7% to 102% across specifications, while the Sustainalytics ENV factor spans from –58% to 6.3%. This means that depending on portfolio construction choices alone, the same underlying rating signal can produce vastly different outcomes.[42]

Consistent with these findings, adding a rating factor to the Fama and French (2018) six-factor (FF6) base portfolio produces, at best, very modest improvements in maximum ex-post Sharpe ratios, with median gains below 0.05 across all providers. Bias-adjusted squared Sharpe ratio differences are similarly negligible and statistically insignificant.[43]

Overall, no robust evidence is found that sustainability ratings systematically predict superior financial performance. The results are not uniformly negative or neutral for all providers and rating types, but the overall picture does not support the notion that higher ratings generally lead to better risk-adjusted returns.

This raises the related, second question: do rating factors carry independent information beyond what is already captured by established asset pricing factors? Factor spanning tests, which regress each rating factor on the FF6 model, show substantial heterogeneity across providers and rating types. The explanatory power of the FF6 factors varies considerably, and most rating factors are not uniformly or strongly spanned, suggesting that they may carry some incremental information.[44]

Whether this information improves the pricing of the cross-section of returns is tested through GRS tests, which assess whether adding a rating factor to the FF6 model reduces the number of jointly significant portfolio alphas. Table 7 reports the percentage-point change in GRS-test rejections at the 5% level when a rating factor is added.

[41] See Paper 4, Appendix D for the full sensitivity of median factor returns to each specification decision.
[42] Cumulative returns are computed over each provider's full sample period: July 2003–June 2024 for Refinitiv and July 2010–June 2024 for Sustainalytics. See Paper 4, Appendix E for the time-series evolution of factor returns and Appendix F for the full distribution of cumulative returns across providers and rating types.
[43] See Paper 4, Figure 6.
[44] See Paper 4, Table 1.

Table 7. GRS-test rejections when adding a sustainability rating factor

	FF6 with A Rating Factor – FF6 %Δ (Δ) p(GRS-Test) <5	
Provider	Sustainability Rating Test Sets	Standard Test Sets
Panel A: Refinitiv		
ESG	–4.5% (–46)	–0.5% (–1)
ENV	–31.2% (–319)	2% (4)
SOC	4.2% (43)	0% (0)
GOV	–8.8% (–90)	0.5% (1)
Panel B: MSCI		
ESG	4.7% (48)	–1.5% (–3)
ENV	–2% (–20)	–1.5% (–3)
SOC	–1% (–10)	–1.5% (–3)
GOV	–2% (–20)	–1% (–2)
Panel C: Bloomberg		
ESG	1% (10)	5% (10)
ENV	–0.5% (–5)	2% (4)
SOC	0.6% (6)	2% (4)
GOV	–1.2% (–12)	2% (4)
Panel D: RobecoSAM		
ESG	0.1% (1)	0% (0)
ENV	–0.1% (–1)	0% (0)
SOC	0.7% (7)	–0.5% (–1)
GOV	–2.1% (–21)	0.5% (1)
Panel E: Sustainalytics		
ESG	–3.4% (–35)	1% (2)
ENV	–7.4% (–76)	2.5% (5)
SOC	–8.4% (–86)	3.5% (7)
GOV	–11.9% (–122)	0% (0)
Panel F: Rating Agreement		
ESG	–2.1% (–22)	4.5% (9)
ENV	0.5% (5)	2% (4)
SOC	2.6% (27)	5% (10)
GOV	–0.4% (–4)	2.5% (5)
Panel G: Rating Disagreement		
ESG	0.5% (5)	0.5% (1)
ENV	–0.2% (–2)	1.5% (3)
SOC	4.9% (50)	0% (0)
GOV	–2.9% (–30)	0.5% (1)

Notes: This table reports the percentage-point change (%Δ) and absolute change (Δ) in the number of GRS-test rejections at the 5% significance level when a single rating factor is added to the Fama-French six-factor (FF6) model, relative to the FF6 baseline. Sustainability Rating Test Sets are matched to the respective rating dimension; Standard Test Sets comprise anomaly portfolios from Hou et al. (2020) and French's data library. Negative values indicate fewer rejections (improved model fit). Sample periods vary by provider availability: Refinitiv 2003–2024, MSCI 2008–2024, Bloomberg 2016–2024, RobecoSAM 2017–2024, Sustainalytics 2010–2024, Rating Agreement/Disagreement 2010–2024.
Source: Paper 4, Table 2.

For the sustainability rating test sets, meaningful improvements are rare. Notable reductions in GRS-test rejections occur only for a few provider-rating combinations, most prominently Refinitiv ENV (−31.2%), Refinitiv GOV (−8.8%), and Sustainalytics GOV (−11.9%). Beyond these isolated cases, the changes are small, generally below 5% in absolute terms and often hovering around zero. For standard test sets, the picture is even clearer with changes remaining negligible across all providers and rating types. Taken together, adding a rating factor to the FF6 model yields only sporadic improvements for selected scores of two providers at best, while offering no observable benefit for explaining the broader cross-section of returns.

Drawing together the findings of Paper 4, no robust evidence is found that higher-rated portfolios or factors systematically generate superior risk-adjusted returns (Question 4a). While isolated positive results emerge for individual providers, most notably Refinitiv and to some extent MSCI, these are not consistent across rating categories, providers, or specification choices. On the asset pricing side, adding a sustainability rating factor to the Fama-French six-factor model does not meaningfully improve its ability to explain the cross-section of returns (Question 4b). Factor spanning tests suggest that rating factors are not fully redundant with established factors, yet GRS test analyses indicate that this incremental information does not translate into improved model fit. Sustainability ratings, despite their central role in investment practice, do not appear to constitute a robust and systematic signal for either portfolio performance or asset pricing.

Finally, this paper is also subject to some limitations. The analysis is restricted to U.S. equities, and the findings may not generalize to other markets where sustainability considerations carry different regulatory or investor weight. Provider and stock coverage varies over time, with some providers entering the sample only in the mid-2010s, which limits the statistical power for shorter-history providers such as Bloomberg and RobecoSAM. Furthermore, while the multi-specification framework substantially broadens the range of construction choices, the space of plausible specification decisions is inherently large, and additional nodes could be explored in future work.

3 Conclusion

3.1 Synthesis of Findings

This dissertation set out to contribute to two fundamental questions in the sustainable finance literature, whether sustainable equity investing strategies affect financial performance, and whether they deliver on their non-financial sustainability objectives. Analogous to Table 2, which mapped each paper onto these two questions, Table 8 provides a high-level summary of the key findings.

Table 8. Summary of key findings across dissertation papers

Paper	Implementation	Financial Performance	Sustainability Effectiveness
1	Commercial sustainable index	Positive, significant	—
2	Commercial sustainable index	—	Lower emissions (−53% to −77%) Higher ratings (varies by provider)
3	Optimized portfolio	(Comparable to or above benchmark)	Large CO_2 reductions (−61% to −93%) Ratings maintained or improved
4	Rating-sorted portfolio	Weak and mixed, provider-dependent	—

Notes: Financial performance for Paper 3 is reported in parentheses as it is a secondary finding. Emission reductions refer to Scope 1+2 carbon emissions relative to the respective benchmark. Paper 2 measures emissions as carbon footprint (emissions per market value), while Paper 3 reports carbon intensity (emissions per revenue) as its primary metric. The ratings improvements (ESG and Environmental) in Paper 2 vary by index variant and rating provider. See Section 2.1 for details.

On the financial performance question, Papers 1 and 3 both find that sustainable investing strategies do not systematically underperform or sacrifice returns. In the case of the sustainable indices examined in Paper 1, all four sustainable variants outperform their standard counterparts on average over the 2015–2023 period. Additionally, the regressions confirm that this outperformance is statistically significant and related to the sustainable index construction criteria rather than compositional differences. Paper 3 supports these findings, as the actively optimized low-carbon portfolios of European equities generate returns that are comparable to or slightly above the benchmark, with Sharpe ratios increasing modestly from 0.61 to 0.73 depending on the tracking error budget. The concern implied by traditional portfolio theory that restricting the investable universe may lead to suboptimal allocations is not supported by the results of either paper.

Paper 4, however, introduces an important qualification. When sustainability ratings are examined directly as investment signals across five major providers, four sustainability ratings, and 512 portfolio construction specifications, the analysis finds no robust or systematic relationship between higher ratings and superior risk-adjusted returns. Isolated positive results appear for individual providers, but these do not generalize across rating categories, providers, or specification choices. Furthermore, influential findings from the prior literature appear fragile when subjected to multi-specification analysis and extended sample periods. Additionally, extending the Fama-French six-factor model with a sustainability rating factor results in no meaningful improvement in explaining the cross-section of returns.

On the sustainability effectiveness question, Paper 2 demonstrates that all four MSCI sustainable index variants achieve meaningfully lower carbon emissions than their standard counterparts, with reductions ranging from 53% for the Climate Change index to 77% for the Paris-Aligned Benchmark over the study period. As with the financial performance results, the regressions also confirm that these reductions are statistically significant and linked to the construction criteria. Paper 3 extends these findings by showing that portfolio optimization can push Scope 1+2 carbon intensity reductions to 61% at a tracking error of just 0.64%, and to over 93% when larger deviations from the benchmark are allowed. The paper also shows that within-sector stock selection, rather than sector tilts, accounts for the dominant share of emission reductions, which suggests that decarbonization need not come at the cost of significant sector bets.

The results for broader sustainability measures, however, are less conclusive. While the SRI and SRI PAB indices consistently achieve the highest improvements in environmental and ESG ratings across providers, the PAB shows limited or even negative effects on these broader measures under Refinitiv and Bloomberg, despite delivering the strongest carbon reductions. The observation that carbon performance and broader sustainability assessment can point in different directions reflects the underlying disagreement among rating providers.

3.2 Limitations and Future Research

The findings of this dissertation are also subject to several limitations that cut across all papers. The most important concerns the sample periods. Papers 1 and 2 cover 2015–2022 (2023 for returns), Paper 3 spans 2011–2024, and Paper 4 covers 2003–2024, depending on provider availability. While Paper 4 has a longer horizon that provides greater statistical power, it also includes the early years during which stock coverage and the number of rating providers were limited. Although all sample periods cover the period of relative macroeconomic stability preceding the COVID-19 pandemic, the pandemic itself, and the war in Ukraine, only Paper 4 includes the global financial crisis and none extend back to the dot-com era. The financial performance results may therefore not generalize to all market environments.

Second, the geographic scope varies across papers but remains limited in each case. Papers 1 and 2 examine global, U.S., European, and Emerging Market equities by analyzing broad regional indices, which may overlook country-specific dynamics. Papers 3 and 4 focus on European and U.S. equities, respectively, and their findings may not transfer to markets with different regulatory environments, investor compositions, or levels of ESG disclosure.

Third, all four papers rely on sustainability data that are subject to well-documented measurement challenges (Berg et al., 2022; Popescu et al., 2023). While each paper addresses these concerns through robustness tests with multiple data sources, the underlying data quality and consistency remain a fundamental challenge for empirical research in sustainable finance more broadly.

Finally, all four papers focus exclusively on equities. While sustainability considerations are becoming increasingly relevant for other asset classes, in particular fixed income for both sovereign and corporate bond investors, the findings of this dissertation cannot be directly extended to these contexts.

These limitations point to several directions for future work. For example, future research could examine how sustainable strategies perform under different macroeconomic regimes and explore country-specific effects in emerging markets where regulatory and disclosure environments vary considerably. Additionally, the analytical framework employed in this dissertation, in particular the multi-specification analysis and the GCE index regressions, could be extended to multi-asset contexts.

3.3 Concluding Remarks

Ultimately, are the results of this dissertation good news or bad news for sustainable investors?

The central finding of this dissertation is that sustainable strategies, whether implemented through commercially available index products or through active portfolio optimization, do not impose a systematic performance penalty and even frequently outperform their conventional counterparts, while delivering measurable sustainability improvements. This suggests that financial and sustainability objectives are not inherently in conflict, at least within the equity asset class and the markets and time periods examined. Even conservative optimization approaches with tight tracking error budgets can achieve carbon intensity reductions above 60%, demonstrating that meaningful decarbonization does not require large departures from established benchmarks. However, the results also caution against overly simplistic expectations. Sustainability ratings do not constitute a robust investment signal, and seemingly small portfolio construction choices can substantially shift empirical outcomes. Moreover, the choice of rating provider matters for both financial and sustainability assessments, and different sustainable strategies can lead to different outcomes. The value proposition of sustainable investing thus depends on the individual investor's objectives and the utility they derive from non-financial outcomes. For investors who seek competitive returns and a lower carbon footprint, the evidence is strongly supportive. For those who expect sustainability ratings to serve as a systematic source of alpha, the evidence is not.

These complexities highlight the importance of methodologically transparent empirical research, to which this dissertation contributes within the sustainable investing literature. The regression framework in Papers 1 and 2 provides an approach for isolating the effect of index construction criteria on both financial and environmental outcomes, rather than simply comparing sustainable and conventional products. The optimization in Paper 3 establishes achievable decarbonization benchmarks without relying on proprietary models. The multi-specification analysis in Paper 4 demonstrates the extent to which portfolio construction choices can drive empirical results in sustainable investing research, a concern that deserves wider recognition in the field.

With sustainable fund assets now exceeding \$3.7 trillion globally (Bioy et al., 2025), sustainable investing will likely continue to grow in scale and importance, and with it the need for rigorous and methodologically transparent empirical evidence.

References

Abouarab, R., T. Mishra, and S. Wolfe. 2025. "Spotting portfolio greenwashing in sustainable funds." *Journal of Business Ethics* 197 (4): 811–839. https://doi.org/10.1007/s10551-024-05783-z.

Alves, R., P. Krüger, and M. 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.

Andersson, M., P. Bolton, and F. Samama. 2016. "Hedging climate risk." *Financial Analysts Journal* 72 (3): 13–32. https://doi.org/10.2469/faj.v72.n3.4.

Atz, U., T. Van Holt, Z. Z. Liu, and C. 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, D., S. Cheng, A. Lioui, and A. Tarelli. 2022. "Sustainable investing with ESG rating uncertainty." *Journal of Financial Economics* 145 (2, Part B): 642–664. https://doi.org/10.1016/j.jfineco.2021.09.009.

Bajic, A., R. Kiesel, and M. Hellmich. 2023. "Handle with care: Challenges in company-level emissions data for assessing financial risks from climate change." *Journal of Climate Finance* 5: 100017. https://doi.org/10.1016/j.jclimf.2023.100017.

Balp, G., and G. Strampelli. 2022. "Institutional investor ESG engagement: The European experience." *European Business Organization Law Review* 23 (4): 869–904. https://doi.org/10.1007/s40804-022-00266-y.

Bell, M., and B. Taylor. 2024. "How can investors balance short-term demands with long-term value?" EY Insights, December 10, 2024. Accessed 20 July, 2025. https://web.archive.org/web/20250725165911/https://www.ey.com/en_gl/insights/climate-change-sustainability-services/institutional-investor-survey.

Bender, J., C. He, C. Ooi, and X. Sun. 2020. "Reducing the carbon intensity of low volatility portfolios." *The Journal of Portfolio Management* 46 (3): 108–122. https://doi.org/10.3905/jpm.2020.46.3.108.

Bender, J., T. A. Bridges, and K. Shah. 2019. "Reinventing climate investing: Building equity portfolios for climate risk mitigation and adaptation." *Journal of Sustainable Finance & Investment* 9 (3): 191–213. https://doi.org/10.1080/20430795.2019.1579512.

Berg, F., J. F. Kölbel, and R. Rigobon. 2022. "Aggregate confusion: The divergence of ESG ratings." *Review of Finance* 26 (6): 1315–1344. https://doi.org/10.1093/rof/rfac033.

Beyer, V., and T. Bauckloh. 2024. "Non-standard errors in carbon premia." CFR Working Paper Series 24-06, Centre for Financial Research (CFR), University of Cologne. https://www.cfr-cologne.de/download/workingpaper/cfr-24-06.pdf.

Bioy, H., B. Wang, A. Lennkvist, L. Mitchell, S. Popat, H. Beaudoin, and J. Yanagi. 2025. "Global sustainable fund flows: Q3 2025 in review." Morningstar Sustainalytics. Accessed January 29, 2026. https://web.archive.org/web/20260114163136/https://www.morningstar.com/content/cs-assets/v3/assets/blt9415ea4cc4157833/blt0ed076e5482cacec/68fa4f8b0e43824cd6c2e60b/Global_ESG_Q3_2025_Flow_Report_October_2025.pdf.

Blitz, D., M. Chen, C. Howard, and H. Lohre. 2024. "3D investing: Jointly optimizing return, risk, and sustainability." *Financial Analysts Journal* 80 (3): 59–75. https://doi.org/10.1080/0015198X.2024.2335142.

Bolognesi, E., A. Dreassi, M. Migliavacca, and A. Paltrinieri. 2024. "Mapping sustainable investing: Exploring ambiguities and consistencies among sustainable indices." *Journal of Environmental Management* 367: 122081. https://doi.org/10.1016/j.jenvman.2024.122081.

Bolton, P., M. Kacperczyk, and F. Samama. 2022. "Net-zero carbon portfolio alignment." *Financial Analysts Journal* 78 (2): 19–33. https://doi.org/10.1080/0015198X.2022.2033105.

Bouchey, P., M. de Leon, Z. Jawaid, and V. Nemtchinov. 2024. "Reducing the carbon footprint of an index: How low can you go?" *The Journal of Wealth Management* 27 (2): 57–74. https://doi.org/10.3905/jwm.2024.1.245.

Brinson, G. P., L. R. Hood, and G. L. Beebower. 1986. "Determinants of portfolio performance." *Financial Analysts Journal* 42 (4): 39–44. https://doi.org/10.2469/faj.v42.n4.39.

Busch, T., M. Johnson, and T. Pioch. 2022. "Corporate carbon performance data: Quo vadis?" *Journal of Industrial Ecology* 26 (1): 350–363. https://doi.org/10.1111/jiec.13008.

Cao, S., Y. Chen, H. Xiao, and A. L. Zhang. 2026. "The economics of greenwashing funds." SSRN Working Paper. https://ssrn.com/abstract=5040390.

Carhart, M. M. 1997. "On persistence in mutual fund performance." *The Journal of Finance* 52 (1): 57–82. https://doi.org/10.1111/j.1540-6261.1997.tb03808.x.

Cenedese, G., S. Han, and M. T. Kacperczyk. 2023. "Carbon-transition risk and net-zero portfolios." SSRN Working Paper. https://doi.org/10.2139/ssrn.4565220.

Chalissery, N., M. I. Tabash, M. Nishad T, and M. A. Saleh Al-Faryan. 2023. "A bibliometric analysis of socially responsible investment based on thematic clustering." *Cogent Business & Management* 10 (1): 2154057. https://doi.org/10.1080/23311975.2022.2154057.

Ciciretti, R., A. Dalò, and L. 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.

Copernicus Climate Change Service. 2025. "Copernicus: 2024 is the first year to exceed 1.5°C above pre-industrial level." Press release, January 10, 2025. Accessed January 29, 2026. https://climate.copernicus.eu/copernicus-2024-first-year-exceed-15degc-above-pre-industrial-level.

Cunha, F. A. F. D. S., E. M. De Oliveira, R. J. Orsato, M. C. Klotzle, F. L. Cyrino Oliveira, and R. G. G. Caiado. 2020. "Can sustainable investments outperform traditional benchmarks? Evidence from global stock markets." *Business Strategy and the Environment* 29 (2): 682–697. https://doi.org/10.1002/bse.2397.

D'Ecclesia, R. L., G. Morelli, and K. Stefanelli. 2022. *Are the green ETFs really green?* Rapporto Tecnico No. 4. Dipartimento di Scienze Statistiche, Sapienza Università di Roma. Accessed January 31, 2026. https://www.dss.uniroma1.it/it/system/files/pubblicazioni/Stefanelli%2C%20Morelli%2C%20D%27Ecclesia%202.pdf.

Delmas, M. A., and V. C. Burbano. 2011. "The drivers of greenwashing." *California Management Review* 54 (1): 64–87. https://doi.org/10.1525/cmr.2011.54.1.64.

Díaz, V., D. Ibrushi, and J. 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.

Dimson, E., O. Karakaş, and X. Li. 2015. "Active ownership." The Review of Financial Studies 28 (12): 3225–3268. https://doi.org/10.1093/rfs/hhv044.

Dobrick, J., C. Klein, and B. Zwergel. 2025. "ESG as risk factor." *Journal of Asset Management* 26 (1): 44–70. https://doi.org/10.1057/s41260-024-00382-z.

Dumitrescu, A., J. Gil-Bazo, and F. Zhou. 2025. "Defining greenwashing." SSRN Working Paper. Accessed January 31, 2026. https://doi.org/10.2139/ssrn.4098411.

Dumitrescu, A., J. Järvinen, and M. Zakriya. 2023. "Hidden gem or fool's gold: Can passive ESG ETFs outperform the benchmarks?" *International Review of Financial Analysis* 86: 102540. https://doi.org/10.1016/j.irfa.2023.102540.

European Commission. 2020. Commission Delegated Regulation (EU) 2020/1818 of 17 July 2020 supplementing Regulation (EU) 2016/1011 of the European Parliament and of the Council as regards minimum standards for EU Climate Transition Benchmarks and EU Paris-aligned Benchmarks. *Official Journal of the European Union* L 406: 17–25. Available at EUR-Lex (CELEX: 32020R1818). Accessed January 30, 2026. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX%3A32020R1818.

European Sustainable Investment Forum. 2016. *European SRI study 2016.* Accessed January 30, 2026. https://www.eurosif.org/wp-content/uploads/2022/03/Eurosif-SRI-study-2016.pdf.

Fahlenbrach, R., and E. Jondeau. 2023. "Greening the Swiss National Bank's portfolio." *The Review of Corporate Finance Studies* 12 (4): 792–833. https://doi.org/10.1093/rcfs/cfad011.

Fama, E. F. 1970. "Efficient capital markets: A review of theory and empirical work." The Journal of Finance 25 (2): 383–417. https://doi.org/10.1111/j.1540-6261.1970.tb00518.x.

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

Fiordelisi, F., G. Galloppo, G. Lattanzio, and V. Paimanova. 2023. "Looking at socially responsible investment strategies through the lenses of the global ETF industry." *Journal of International Money and Finance* 137: 102917. https://doi.org/10.1016/j.jimonfin.2023.102917.

Friede, G., T. Busch, and A. 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.

FTSE Russell. 2022. *Asset owners widely adopting sustainable investment: 2022 global survey findings from asset owners*. FTSE Russell. Accessed January 29, 2026. https://www.lseg.com/content/dam/ftse-russell/en_us/documents/reports/global-asset-owner/sustainable-investment-2022-global-survey-findings-from-asset-owners.pdf.

FTSE Russell. 2025. *FTSE Russell ESG data model methodology*. v1.1, August 2025. Accessed February 1, 2026. https://www.lseg.com/content/dam/ftse-russell/en_us/documents/policy-documents/ftse-russell-esg-data-model-methodology.pdf.

Gaskell, N. 2025. "Why the choice of carbon metrics matters." Aberdeen Investments. September 10, 2025. Accessed February 1, 2026. https://www.aberdeeninvestments.com/en-gb/intermediary/insights-and-research/why-the-choice-of-carbon-metrics-matters.

Gazali, S., S. Aziz, and M. I. Nazir. 2025. "Socially responsible investment (SRI) a scientometric and bibliometric analysis and future research agenda (1991–2023)." *Environment, Development and Sustainability* (2025). https://doi.org/10.1007/s10668-025-06096-5.

Gibbons, M. R., S. A. Ross, and J. Shanken. 1989. "A test of the efficiency of a given portfolio." *Econometrica* 57 (5): 1121–1152. https://doi.org/10.2307/1913625.

Gibson-Brandon, R., P. Krueger, and P. S. Schmidt. 2021. "ESG rating disagreement and stock returns." *Financial Analysts Journal* 77 (4): 104–127. https://doi.org/10.1080/0015198X.2021.1963186.

Giglio, S., M. Maggiori, J. Stroebel, Z. Tan, S. P. Utkus, and X. Xu. 2025. "Four facts about ESG beliefs and investor portfolios." *Journal of Financial Economics* 164: 103984. https://doi.org/10.1016/j.jfineco.2024.103984.

Global Impact Investing Network. 2025. "What you need to know about impact investing." January 24, 2025. Accessed February 14, 2026. https://thegiin.org/publication/post/about-impact-investing/.

Global Sustainable Investment Alliance. 2025. *Global sustainable investment review 2024.* Accessed January 30, 2026. https://www.gsi-alliance.org/wp-content/uploads/2025/11/GSIR-2024-Main-Report.pdf.

Gutsche, G., H. Wetzel, and A. Ziegler. 2023. "Determinants of individual sustainable investment behavior - A framed field experiment." *Journal of Economic Behavior & Organization* 209: 491–508. https://doi.org/10.1016/j.jebo.2023.03.016.

Halbritter, G., and G. 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.

Heeb, F., J. F. Kölbel, F. Paetzold, and S. Zeisberger. 2023. "Do investors care about impact?" *The Review of Financial Studies* 36 (5): 1737–1787. https://doi.org/10.1093/rfs/hhac066.

Heldmann, J., H. D. Dang, and T. Brückner. 2025. "Investing in a low-carbon transition: Carbon footprint savings with green MSCI indices." *Journal of Environmental Management* 393: 126022. https://doi.org/10.1016/j.jenvman.2025.126022.

Henriquez-Salman, R. 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, K., C. Xue, and L. Zhang. 2020. "Replicating anomalies." *The Review of Financial Studies* 33 (5): 2019–2133. https://doi.org/10.1093/rfs/hhy131.

Hübel, B., and H. 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.

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, 2023. 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/.

Jacob, A., and M. Wilkens. 2021. "What drives sustainable indices? A framework for analyzing the sustainable index landscape." SSRN Working Paper. https://doi.org/10.2139/ssrn.3874437.

Jain, M., G. D. Sharma, and M. Srivastava. 2019. "Can sustainable investment yield better financial returns: A comparative study of ESG indices and MSCI indices." *Risks* 7 (1): 15. https://doi.org/10.3390/risks7010015.

Jin, I. 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.

Jondeau, E., B. Mojon, and L. A. Pereira da Silva. 2025. "Building benchmark portfolios with decreasing carbon footprints." *Journal of Sustainable Finance & Investment*. Advance online publication. https://doi.org/10.1080/20430795.2025.2463434.

Jonwall, R., S. Gupta, and S. Pahuja. 2024. "Do socially responsible indices outperform conventional indices? Evidence from before and after the onset of COVID-19." *Corporate Social Responsibility and Environmental Management* 31 (5): 4995–5011. https://doi.org/10.1002/csr.2839.

Kaustia, M., and W. Yu. 2021. "Greenwashing in mutual funds." SSRN Working Paper. https://doi.org/10.2139/ssrn.3934004.

Khan, M. 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.

Kolasa, T., and Z. Sautner. 2025. “Institutional investors and the fight against climate change.” *Corporate Governance: An International Review* 33 (4): 663–679. https://doi.org/10.1111/corg.12620.

Krueger, P., Z. Sautner, and L. T. Starks. 2020. “The importance of climate risks for institutional investors.” *The Review of Financial Studies* 33 (3): 1067–1111. https://doi.org/10.1093/rfs/hhz137.

Lean, H. H., and D. K. Nguyen. 2014. “Policy uncertainty and performance characteristics of sustainable investments across regions around the global financial crisis.” *Applied Financial Economics* 24 (21): 1367–1373. https://doi.org/10.1080/09603107.2014.925063.

Lee, L.-E., T. Séné, K. Towey, R. Mahmood, and U. Ashfaq. 2026. *Transition finance tracker: Fourth quarter 2025*. MSCI Institute. Accessed January 29, 2026. https://www.msci-institute.com/wp-content/uploads/2026/01/MSCI-Transition-Finance-Tracker-Q4-2025-290125_2.pdf.

Lintner, J. 1965. “The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets.” *The Review of Economics and Statistics* 47 (1): 13–37. https://doi.org/10.2307/1924119.

LSEG. 2024a. *Environmental, social and governance scores from LSEG*. October 2024. Accessed February 1, 2026. https://www.lseg.com/content/dam/data-analytics/en_us/documents/methodology/lseg-esg-scores-methodology.pdf.

LSEG. 2024b. *LSEG ESG carbon data and estimate models.* Fact sheet. Accessed February 2, 2026. https://www.lseg.com/content/dam/data-analytics/en_us/documents/fact-sheets/lseg-esg-carbon-data-and-estimate-models.pdf.

Lublóy, Á., J. L. Keresztúri, and E. Berlinger. 2025. “Quantifying firm-level greenwashing: a systematic literature review.” *Journal of Environmental Management* 373: 123399. https://doi.org/10.1016/j.jenvman.2024.123399.

Luo, W., Z. Tian, S. Zhong, Q. Lyu, and M. Deng. 2022. “Global evolution of research on sustainable finance from 2000 to 2021: A bibliometric analysis on WoS database.” *Sustainability* 14 (15): 9435. https://doi.org/10.3390/su14159435.

Madhavan, A., A. Sobczyk, and A. 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.

Margolis, J. D., H. A. Elfenbein, and J. 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 Working Paper. https://doi.org/10.2139/ssrn.1866371.

Markowitz, H. 1952. “Portfolio selection.” *The Journal of Finance* 7 (1): 77–91. https://doi.org/10.1111/j.1540-6261.1952.tb01525.x.

Markowitz, H. 1959. *Portfolio selection: Efficient diversification of investments.* New York: John Wiley & Sons.

Mazzolini, G., D. Salakhova, M. Giuzio, and S. Kapadia. 2025. "Sustainability labels vs. reality: how climate-friendly are green and ESG funds?" *Working Paper Series* No. 3121. European Central Bank, Frankfurt am Main, Germany. Accessed January 29, 2026. https://www.ecb.europa.eu/pub/pdf/scpwps/ecb.wp3121~47d80e073f.en.pdf.

Meyers, M. S., I. Ferrero-Ferrero, and M. J. Muñoz-Torres. 2024. "Are sustainable funds doing the talk and the walk? An ESG score analysis of fund portfolio holdings." *International Review of Economics & Finance* 93 (part A): 1526–1541. https://doi.org/10.1016/j.iref.2024.04.023.

Morgan Stanley Institute for Sustainable Investing. 2025. *Sustainable signals: Individual investors 2025*. Morgan Stanley Institute for Sustainable Investing. Accessed January 29, 2026. https://www.morganstanley.com/content/dam/msdotcom/en/assets/pdfs/2025_Sustainable_Signals_Individual_Investors_2025_report.pdf.

Mossin, J. 1966. "Equilibrium in a capital asset market." *Econometrica* 34 (4): 768–783. https://doi.org/10.2307/1910098.

MSCI. 2024a. *MSCI SRI Filtered PAB Indexes methodology.* November 2024. Accessed January 31, 2026. https://www.msci.com/documents/10199/d21129da-4fe6-9c55-81d0-eb42d4931b51.

MSCI. 2024b. *ESG ratings methodology*. April 2024. Accessed February 1, 2026. https://www.msci.com/documents/1296102/34424357/MSCI%2BESG%2BRatings%2BMethodology.pdf.

MSCI. 2024c. *Greenhouse gas emissions methodologies.* MSCI ESG Research LLC. June 2024. Accessed February 2, 2026. https://www.msci.com/documents/1296102/39141520/MSCI%2BGreenhouse%2BGas%2BEmissions%2BMethodologies.pdf.

Munich Re. 2025. *Climate change is showing its claws: The world is getting hotter, resulting in severe hurricanes, thunderstorms and floods.* Munich Re. January 9, 2025. Accessed January 29, 2026. https://www.munichre.com/en/company/media-relations/media-information-and-corporate-news/media-information/2025/natural-disaster-figures-2024.html.

Naffa, H., and M. 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.

Otchere, I., A. Abdulrahman, and J. Wang. 2025. "Environmental, social, and governance investing and sustainability of pension funds: Evidence from the organisation for economic cooperation and development countries." *Economic Modelling* 143: 106948. https://doi.org/10.1016/j.econmod.2024.106948.

Pástor, Ľ., R. F. Stambaugh, and L. 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, Ľ., R. F. Stambaugh, and L. A. Taylor. 2022. "Dissecting green returns." *Journal of Financial Economics* 146 (2): 403–424. https://doi.org/10.1016/j.jfineco.2022.07.007.

Pavlova, I., and M. E. de Boyrie. 2022. "ESG ETFs and the COVID-19 stock market crash of 2020: Did clean funds fare better?" *Finance Research Letters* 44: 102051. https://doi.org/10.1016/j.frl.2021.102051.

Pedersen, L. H., S. Fitzgibbons, and L. 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.

Popescu, I.-S., T. Gibon, C. Hitaj, M. Rubin, and E. Benetto. 2023. "Are SRI funds financing carbon emissions? An input-output life cycle assessment of investment funds." *Ecological Economics* 212: 107918. https://doi.org/10.1016/j.ecolecon.2023.107918.

Riedl, A., and P. Smeets. 2017. "Why do investors hold socially responsible mutual funds?" *The Journal of Finance* 72 (6): 2505–2550. https://doi.org/10.1111/jofi.12547.

Robeco. 2025. *2025 global climate investing survey: Balancing risk, return and sustainability in turbulent times*. https://www.robeco.com/files/docm/docu-202506-robeco-global-climate-investing-survey-2025.pdf.

Rompotis, G. G. 2023. "The performance of ESG ETFs in the U.S." *Capital Markets Review* 31 (2): 89–101. https://www.mfa.com.my/wp-content/uploads/2023/09/v31_i2_a5_pg89-101.pdf.

Schoenmaker, D. and W. Schramade. 2019. *Principles of sustainable finance*. Oxford: Oxford University Press.

Schueth, S. 2003. "Socially responsible investing in the United States." *Journal of Business Ethics* 43 (3): 189–194. https://doi.org/10.1023/A:1022981828869.

Setzer, J., and C. Higham. 2025. *Global trends in climate change litigation: 2025 snapshot*. London: Grantham Research Institute on Climate Change and the Environment, London School of Economics and Political Science. https://doi.org/10.21953/LSE.LH46LE9Y8SGI.

Sharpe, W. F. 1964. "Capital asset prices: a theory of market equilibrium under conditions of risk." *The Journal of Finance* 19 (3): 425–442. https://doi.org/10.1111/j.1540-6261.1964.tb02865.x.

Shemfe, M., E. Bourne, A. Meng, J. Simmons, F. Fouret, M. Jain, and J. Kooroshy. 2024. *Decarbonisation in portfolio benchmarks: Tracking portfolio carbon transition.* FTSE Russell, London Stock Exchange Group. September 2024. Accessed February 1, 2026. https://www.lseg.com/content/dam/lseg/en_us/documents/sustainability/lseg-decarbonisation-portfolio-benchmarks-report.pdf.

Sherwood, M. W., and J. Pollard. 2023. *Responsible investing: an introduction to environmental, social, and governance investments*. 2nd ed. Abingdon, Oxon and New York: Routledge. https://doi.org/10.4324/9781003213666.

Siemroth, C., and L. Hornuf. 2023. "Why do retail investors pick green investments? A lab-in-the-field experiment with crowdfunders." *Journal of Economic Behavior & Organization* 209: 74–90. https://doi.org/10.1016/j.jebo.2023.02.023.

Soebhag, A., B. Van Vliet, and P. Verwijmeren. 2024. "Non-standard errors in asset pricing: mind your sorts." *Journal of Empirical Finance* 78: 101517. https://doi.org/10.1016/j.jempfin.2024.101517.

Swiss Re. 2024. *sigma Resilience Index 2024: Encouraging resilience gains, but more is needed.* Swiss Re Institute. July 24, 2024. Accessed January 29, 2026. https://www.swissre.com/institute/research/sigma-research/natural-catastrophe-insurance-global-resilience-index-2024.html.

Walter, D., R. Weber, and P. Weiss. 2024. "Methodological uncertainty in portfolio sorts." SSRN Working Paper. https://doi.org/10.2139/ssrn.4164117.

Whelan, T., U. Atz, T. Van Holt, and C. Clark. 2021. *ESG and financial performance: Uncovering the relationship by aggregating evidence from 1,000 plus studies published between 2015-2020.* NYU Stern Center for Sustainable Business and Rockefeller Asset Management. Accessed January 30, 2026. https://www.stern.nyu.edu/sites/default/files/assets/documents/NYU-RAM_ESG-Paper_2021%20Rev_0.pdf.

World Meteorological Organization. 2025. "WMO confirms 2024 as warmest year on record at about 1.55°C above pre-industrial level." Press Release, January 10, 2025. Accessed January 29, 2026. https://wmo.int/news/media-centre/wmo-confirms-2024-warmest-year-record-about-155degc-above-pre-industrial-level.

Xiao, Y., R. Faff, P. Gharghori, and D. 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.

Appendix

Each paper is presented as a self-contained work with its own footnotes, references, and appendices. Papers 1 and 2 appear in their published form, while Papers 3 and 4 reflect their most recent working versions. Differences in formatting, spelling conventions, and reference styles across the four papers are attributable to the conventions and requirements of the respective target journals.

Paper 1 – Financial Returns of Going Green: Evidence from MSCI Indices

Abstract*

This study examines how the green criteria (GCE) used by MSCI to create green equity indices influence their financial performance. We analyze the Climate Change (CC), Paris-Aligned Benchmark (PAB), Socially Responsible Investment (SRI), and SRI Filtered PAB (SRI PAB) index variants in comparison with their standard non-green counterpart in each of the four regions: the World, the USA, Europe, and Emerging Markets (EM). Overall, the green indices often matched or exceeded the returns of their standard index without adding significant risk. With few exceptions in the EM, the green indices exhibited better long-term financial performance than their standard index. Over 2015–2023, the CC, PAB, SRI, and SRI PAB indices respectively delivered cumulative excess returns of 4.7%, 5.8%, 13.7%, and 7.5% relative to the standard index. Their returns co-moved closely with the market and the standard index's returns. The GCEs statistically and significantly contributed to green indices' relative financial outperformance.

Keywords: MSCI; Green index; Cumulative return differential; Wealth relative

* This paper is joint work with Thomas Brückner and Huong Dang. It has been published as: Heldmann, J., T. Brückner, and H. D. Dang. 2025. "Financial returns of going green: evidence from MSCI indices." *Journal of Asset Management* 26: 768–787. https://doi.org/10.1057/s41260-025-00404-4. This version includes minor typographical and formatting adjustments. Section numbers are added to facilitate navigation. The content remains unchanged.

1 Introduction

The growing need for investment solutions addressing climate, social, and governance concerns has drawn great investors' interest in green products.[1] At the end of 2023, global sustainable funds managed about US$2.2 trillion and continued to outgrow conventional funds in terms of inflows (Cortina and Phyu 2024). Many of these products use green indices replicated through Exchange Traded Funds (ETFs); however, it remains uncertain whether such green indices/ ETFs consistently outperform their conventional non-green counterparts. Some studies report positive relative financial performance (Fiordelisi et al. 2023; Rompotis 2023; Pavlova and de Boyrie 2022), greater resilience during financial crises (Ortas et al. 2014) and superior performance during the COVID-19 pandemic (Lin and Swain 2024). Conversely, other research finds green indices/ETFs perform comparably to traditional benchmarks, vary across regions, or show no significant difference in risk-adjusted returns (Bolognesi et al. 2024; Jonwall et al. 2024; Dumitrescu et al. 2023; Cunha et al. 2020; Jain et al. 2019; Benson et al. 2010). Some studies even suggest underperformance of green indices/ETFs relative to traditional financial products (Lean and Nguyen 2014; Ortas et al. 2012). Evidence on return correlation is also mixed, with some authors observing decoupling (Ang 2015; Lean and Nguyen 2014) and others documenting high correlations or co-movements (Rompotis 2023; Managi et al. 2012). Thus, it remains unclear whether green indices/ ETFs exhibit better financial performance than their nongreen counterparts. Furthermore, very little effort has been made to directly examine if and how criteria employed to construct green indices relate to their financial performance. If green criteria do not affect a green index's performance, any observed outperformance (relative to its standard nongreen index) is likely due to luck. The existing evidence, mostly for ETFs, implying a mixed relationship (Abate et al. 2021; Papathanasiou and Koutsokostas 2024).

Our study is motivated by the mixed reported results on the financial performance of green indices/ETFs and the unexplored (mixed) relationship between green criteria and the financial performance of green indices (ETFs). In this study, we choose to explore Morgan Stanley Capital International (MSCI) indices as they have been widely used as reference indices for green equity ETFs.[2] We address two main research questions: (1) How did MSCI green indices perform financially compared with their respective standard (parent) index? and (2) How did

[1] In this study, we address various sustainable investments, for example, Environmental, Social & Governance (ESG), Social Responsible Investment (SRI), sustainable, ethical, green as green investments.

[2] As of October 2024, assets under management in ETFs linked to MSCI equity indexes reached approximately $1.72 trillion. See: https://ir.msci.com/aum-etfs-linked-msci-indexes.

the green criteria employed by MSCI to create each green index affect its financial performance?

We focus on indices rather than ETFs for several reasons. First, indices are more suitable for performance analysis, as the impact of total expense ratio, taxes,[3] distribution policy,[4] and replication strategy[5] varies across providers and can obscure return analyses. Second, the exact benchmark an ETF tracks can change over time, while the ETF keeps the same International Securities Identification Numbers (ISIN), which can distort benchmark-adjusted return analyses.[6] Third, ETF providers use slightly modified benchmarks regarding exclusion thresholds, compared with our benchmark being the standard non-green index. That is, providers aim to achieve a performance advantage in their respective peer group by offering ETFs that do not track exactly the competitor's benchmark but the same green exposure.[7]

We examine a set of 20 MSCI indices spanning four major regions: the World, the USA, Europe, and emerging markets (EM). Within each region, we investigate the performance of each of the four green indices—the Climate Change (CC), the Paris-Aligned Benchmark (PAB), the Socially Responsible Investment (SRI), and the SRI filtered PAB (SRI PAB) index—relative to the standard (parent) MSCI index. Our primary dataset on index returns extends from January 2015 to December 2023, providing 2160 monthly index observations. Considering the availability of firm-level data and Tobin's Q measure, our sample of 1,076,340 firm-month observations is limited to December 2022 and includes 5267 unique constituents of 50 countries and 11 sectors. The rich historical data of MSCI index constituents enables us to carry out more thorough and reliable analyses than using comparable ETFs.[8]

[3] Withholding tax benefits on US dividends, depending on the fund domicile of an ETF, can have a substantial impact on fund performance. Physically replicating ETFs with US exposure domiciled in Ireland have a reduced withholding tax burden of 15% compared to, for example, 30% for ETFs domiciled in Luxembourg.
[4] For example, accumulating ETFs reinvest their dividends while distributing ETFs pay out dividends.
[5] There are two main replication strategies, physical and synthetic replication. They split into full or physically optimized replication and swap replication (which can be unfunded, funded and fully funded). The replication strategy directly influences the performance of an ETF. Physical replication benefits from securities lending while those using swaps are, depending on the qualification of the index under Section 871(m) of the US Internal Revenue Code, not burdened by tax regulations. Generally, ESG ETFs use physical replication.
[6] For example, the Amundi Index MSCI Emerging Markets (EM) SRI PAB UCITS ETF (ISIN: LU1861138961) was launched in January 2019 and tracked the MSCI EM SRI Index. In October 2019, the reference index was switched to MSCI EM SRI 5% Issuer Capped. From December 2020 onward, the ETF tracked the MSCI EM SRI Filtered PAB. Another example is the iShares MSCI World SRI UCITS ETF (ISIN: IE00BYX2JD69). It was launched in October 2017 with MSCI World SRI Select Index being the benchmark. In November 2019, it started tracking the MSCI World SRI Select Reduced Fossil Fuel Index.
[7] Please see the green product range (SFDR Article 8 and 9) from Amundi, www.amundietf.nl/en/professional/etf-products/search, iShares, www.ishares.com/us/products/etf-investments and Xtrackers, www.etf.dws.com/en-us/etf-products.
[8] MSCI provides current data on its indices in the factsheets and via the constituent download function at https://www.msci.com/const ituents. Historical data is not freely available, but our data (continued on next page)

Our study differs from recent research on ETFs' financial performance, such as ElBannan (2024). First, we focus on indices instead of ETFs. By doing so, we avoid the added complexity of taxes, distribution policies, and replication methods that can influence ETF returns. Second, our index constituent data allows us to control for differences in sectors and levels of development across countries. Third, we analyze four distinct types of green indices rather than treating all ESG or SRI products as one category. This includes giving special attention to the Paris-Aligned Benchmark (PAB) and SRI PAB indices which both have not been studied in the literature.

Our study is related to Kossentini et al. (2024) and Jacob and Wilkens (2021). The former focus on four MSCI ESG Leader indices covering different parts of the European region: Europe, the Economic and Monetary Union (EMU), Emerging Markets Europe, and the Middle East and Europe, for the years from 2007 to 2020. The latter analyze four ESG and four carbon-focused MSCI World Indices from 2011 to 2019. Out of our 16 green indices, only the World SRI overlaps with Jacob and Wilkens's (2021) sample. We extend these two studies by covering a more recent period that includes significant global events such as the onset of the war in Ukraine. Our study encompasses two additional major regions, the USA, and Emerging Markets. Importantly, we place particular emphasis on indices aligned with the Paris Climate Agreement, including the PAB and SRI PAB indices, which are of particular interest to climateconscious investors. A recent study conducted by Bolognesi et al. (2024) investigates the MSCI SRI and MSCI PAB indices, but its sample is limited to the USA for a short period, August 2021–May 2024. We employ a richer dataset that covers four regions over a longer period (2015–2023). Most importantly, we carry out index-level regression analyses that control for sector- and country-specific parameters. This allows us to directly evaluate the effectiveness of the green criteria (GCE) employed by MSCI to create each green index type.

To assess the financial performance of a green index, we use cumulative return differential (CRD) as our primary measure.[9] For robustness tests, we utilize Wealth Relative (WR), Tobin's Q,[10] the Sharpe ratio, and the Treynor ratio. CRD and WR capture the performance of a green

set extends from the beginning of 2015 to the end of 2023. The index performance of 12 indices launched after 2015 has been reconstructed (by MSCI) using the identical methods and principles of MSCI. This historical data availability makes our study unique, as we are not aware of any other study that analyses MSCI PAB and SRI PAB indices over such a long period of time, including the Corona crisis and the Ukraine war.

[9] As we compare MSCI green indices to their corresponding MSCI standard index, we do not need to consider index provider' characteristics that affect an index's financial performance.

[10] We avoid using accounting figures from individual companies to compute index-level return on equity (ROE) and return on assets (ROA) because several factors can distort these accounting return measures. Differences in accounting standards and legal frameworks across countries can lead to inconsistent financial reporting. In addition, managers often have significant flexibility, which may allow them to adjust or (continued on next page)

index, relative to its standard index, over an investment horizon; each is computed directly at the index level. Other metrics are not relative measures; each is estimated using an index's monthly constituent weights and constituents' metrics. To examine an index's return co-movement with the market index, we estimate the market factor beta coefficient using Fama-French models. We also assess the dynamic conditional correlations between green and standard index returns over time using a GARCH(1,1) model with a Gaussian distribution on daily return data. To explore the relationship between green criteria and a green index's financial performance, we conduct OLS regressions and employ four dummy variables, each represents the impact of MSCI's green criteria (GCE) used to create one green index type.

Our analysis documents strong evidence supporting the attractiveness of green indices. On average, over the period January 2015–December 2023, an investor would have respectively earned a 4.7%, 5.8%, 13.7%, and 7.5% greater cumulative return if s/he had invested in the green CC, PAB, SRI, and SRI PAB index instead of the standard index. Taking into account the expense ratios of the largest ETFs which replicate the green and standard indices, an investor would have respectively earned a net cumulative return differential of 4%, 5.1%, 13%, and 6.8% during 2015–2023. Across the four regions, most green indices demonstrated long-term outperformance (except in EM, where only the SRI outperformed the standard index). On average, green indices across regions achieved greater Sharpe and Treynor ratios and delivered better risk-adjusted returns than their respective standard index. Furthermore, green indices' returns moved closely with those of the market index and their respective standard index. We find strong evidence that MSCI's GCEs statistically and significantly contributed to green indices' better financial performance, relative to the standard index. The positive effects of GCEs were more pronounced on the outperformance of the SRI and SRI PAB indices; both are aligned with the Paris Climate Agreement.

Our main results are consistent across robustness tests, including analyses with one-, two- and three-month lagged control variables, alternative return measures, and various subsamples. Our key finding indicates that green indices can deliver competitive returns and, in most cases, outperform their standard non-green counterparts over a relatively long time period. By emphasizing the role of green criteria in index construction, our study contributes to ongoing discussions on the financial attractiveness of green equity investments.

smooth earnings over time. Furthermore, during periods of high inflation, depreciation expenses may be understated since they do not reflect the true replacement costs of equipment, potentially leading to artificially inflated earnings.

The rest of our paper is structured as follows. The "Brief overview of MSCI's methods for creating green equity indices" section provides a brief overview of the process by which MSCI selects constituents and constructs the examined green indices. The "Literature review and research questions" section reviews the literature and proposes research questions. The "Data and methods" section describes data and presents the method. The "Empirical results" section discusses the empirical results. The "Conclusion" section concludes the key findings.

2 Brief overview of MSCIs method for creating green equity indices

The MSCI Standard (parent) Index serves as the starting universe for determining the eligible universe of a green index.[11] The parent indices in our sample were launched in March 1986, except for EM in January 2001. The green indices are constructed by applying exclusion criteria to the parent index, leading to a smaller eligible universe. The degree of exclusion varies across the green indices, with stricter criteria resulting in fewer components compared to the parent index. Among the indices analyzed, the MSCI Climate Change (CC) Index applies the mildest exclusions, while the MSCI Socially Responsible Investing (SRI) filtered PAB Index enforces the strictest exclusions. Both the PAB and SRI PAB indices are consistent with the Paris Agreement's objectives, with the SRI PAB excluding the most securities.

2.1 The MSCI CC index, and the MSCI climate PAB index

To form the MSCI CC Index (launched in June 2019) and the MSCI Climate PAB Index (launched in October 2020), a negative screening is applied to the parent index in the first step. Companies are excluded from the parent index if they violate either the UN Global Compact Principles or sector policies, such as generating revenue above set thresholds from controversial weapons, tobacco, or fossil fuels. Next, the sustainability of the remaining companies is evaluated using the Low Carbon Transition (LCT) score, which considers three components: a company's carbon footprint, its climate-related risks, and its ability to manage those risks. Companies are categorized into five groups based on their LCT scores: Solutions, Neutral, Operational Transition, Product Transition, and Asset Stranding. Companies classified as Solutions (with the highest LCT score of three) receive a higher weighting in the final index. Weight adjustments are applied across all five categories relative to the security weights in the parent index, forming the final eligible universe for the CC Index. For instance, in the "World" region, this process transforms the MSCI World parent index into the MSCI World CC Index.

[11] See https://www.msci.com/index-methodology for an overview of the construction methodology of all MSCI Indices.

The MSCI Climate PAB Index, however, applies stricter criteria with a PAB overlay aligned with the goals of the Paris Agreement. Specifically, companies must reduce their Weighted Average Carbon Intensity (WACI) by 50% and lower their annual carbon footprint by 7% compared to the parent index. Additionally, the Climate PAB Index excludes companies deriving revenue from fossil fuel activities, such as coal, oil, natural gas exploration, processing, or power generation with excessive greenhouse gas intensity. Finally, the weighting of each company in the MSCI Climate PAB Index is determined using its LCT score.

2.2 The MSCI SRI index, and the MSCI SRI filtered PAB (SRI PAB) index

The MSCI SRI Index (launched in June 2011, except for EM in March 2014) and the MSCI SRI filtered PAB Index (launched in June 2020) apply stricter exclusion criteria than the Climate Indexes discussed above. Both indices implement more sector-specific exclusions with low thresholds, resulting in a significantly smaller eligible universe compared to the parent index. A detailed comparison of the exclusion criteria and thresholds can be found in "Appendix A."

The MSCI SRI filtered PAB Index is more exclusionary than the MSCI SRI Index due to its alignment with the Paris Agreement goals. It applies additional thresholds to companies generating revenue from oil and gas activities. Furthermore, companies with serious violations of sustainable investment objectives are flagged as "Red," receiving an ESG Controversies Score of 0, and are excluded from the index. The remaining companies are rated using MSCI ESG Ratings, which range from AAA to CCC. For inclusion, existing index constituents must have a minimum rating of BB and an ESG Controversies Score of at least 1, while new additions require a minimum rating of A and an ESG Controversies Score of 4.

Both indices employ a best-in-class approach, selecting only the top 25% of companies within each sector. The SRI filtered PAB Index also incorporates the PAB overlay, requiring companies to reduce their WACI by 50% and their carbon footprint by 7% annually relative to the parent index. To ensure diversification, the weight of any single company in the index is capped at 5%.

3 Literature review and research questions

From a portfolio theory perspective, both negative screening and best-in-class approaches used to construct green portfolios limit the investment universe, potentially reducing diversification and increasing risk compared to broader portfolios (Barnett and Salomon 2006; Renneboog et al. 2008). Nevertheless, several studies indicate that green stocks or those with high ESG scores exhibit favorable characteristics, such as lower unsystematic risk (Giese et al. 2019; Hong and

Kacperczyk 2009) and reduced capital costs (Gregory et al. 2021; Unruh et al. 2016; Ng and Rezaee 2015). Most research on green indices or ETFs finds minimal differences in returns compared to benchmarks, and any observed differences are typically aligned with risk-adjusted returns.

Studies on SRI and ESG indices report either better or similar performance compared to conventional benchmarks. Those finding better performance include Statman (2005), Cunha et al. (2020), Jain et al. (2023), and Kossentini et al. (2024). In contrast, studies reporting similar performance include Schröder (2007), Collison et al. (2008), Consolandi et al. (2009), Benson et al. (2010), Jain et al. (2019), Jacob and Wilkens (2021), and Bolognesi et al. (2024). Several studies also document that green indices/ETFs demonstrate greater resilience during crises (Ortas et al. 2014, 2013; Omura et al. 2021; Lin and Swain 2024; Huang 2024; ElBannan 2024). Anti-ESG ETFs, however, tend to underperform (Rompotis 2024).

In terms of returns co-movement, Jain et al. (2019) find that green indices closely track their benchmarks across the USA, E.U., EM, and global markets during 2013–2017. Similarly, Managi et al. (2012) observe no distinct characteristics between green indices and their benchmarks, noting high co-movement across various market conditions. These findings alone suggest that green indices are not markedly distinct investment vehicles compared with their traditional benchmarks.

Despite the growing popularity of green investments, very few empirical studies directly investigate how green indices' construction criteria affect their financial performance. This is likely due to the challenge in obtaining indices' constituent weighting data over a reasonably long period. Related studies mainly examine funds categorized by sustainability/ESG ratings and report mixed findings. Using 634 European mutual funds during 2014–2019, Abate et al. (2021) report better performance for funds with high Morningstar Sustainability Ratings. In contrast, Papathanasiou and Koutsokostas (2024) find that ESG funds with low Morningstar ratings outperform among 235 ESG mutual funds during 2010–2022. Folger-Laronde et al. (2022) observe that higher-rated ETFs with an Eco-Fund Label have lower returns during the COVID-19 market crash. Rompotis (2022a, 2022b) find no significant risk-adjusted return differences for ESG ETFs and a negative relationship between returns and ESG metrics, respectively. Given the above mixed findings, it remains unclear whether sustainability/ESG ratings are linked to ETFs' better financial performance. In reality, the criteria used to construct a green index are complex and extend far beyond ESG ratings, as we briefly discussed in "Brief overview of MSCI's methods for creating green equity indices" section. The unexplored relationship between green criteria and green index's financial performance motivates us to investigate how the green criteria used by MSCI affect its green indices' financial performance.

Our literature review highlights several gaps that have not been well addressed in prior studies. Most studies focus on generic ESG/SRI ETFs/indices without analyzing those aligned with the Paris Agreement's emissions reduction goals. Return comparisons often involve multiple providers against a single benchmark, ignoring providerspecific characteristics. Furthermore, the impact of green criteria used to create an index and its financial performance remains unexplored. Our study addresses these gaps by conducting a comprehensive analysis of green MSCI indices, utilizing constituent weighting data across four major regions, assessing their alignment with the Paris Agreement, and benchmarking them against their respective MSCI standard indices. We aim to address the research gaps by answering the following two research questions:

1a. How did green MSCI indices perform financially compared with its standard non-green index?

1b. How did an index's Green Criteria Effectiveness (GCE) affect its financial performance?

4 Data and methods

4.1 Data

Our study makes use of both index-level and firm-level data. Our monthly index dataset includes 20 MSCI equity indices from January 2015 to December 2023. The five examined indices (one standard and four green-oriented variants) are analyzed in each of the four regions: the World, the USA, EU, and EM. We collect daily Net Total Returns (in USD) from Bloomberg, which include dividend reinvestments and consider withholding taxes. To capture an index's exposure to the market index, we use monthly Fama-French factor returns and risk-free rates from French's (2024) online data library.[12]

For the firm-level analysis, MSCI generously provided us with monthly data on each index's constituents and their weights from January 2015 to December 2022. This dataset includes 5267 unique firms from 11 sectors across 50 countries. We are able to meticulously map 5034 of these firms (covering more than 99.9% of the total indices' weights) to their ISINs. Other firm-specific data is obtained from Refinitiv Eikon's Datastream, which covers nearly our entire sample. To our knowledge, this is one of the first studies that employ such a rich MSCI index data, including two Paris Agreement-aligned indices, over a relatively long period.

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

4.2 Methods

4.2.1 Index's financial performance measures

We employ cumulative return differential (CRD) as our primary measure that captures the financial performance of a green index gi, relative to its parent index pi. The return differential $rd_{gi,t}$ of green index gi relative to its parent index pi at time t is computed as:

$$rd_{gi,t} = r_{gi,t} - r_{pi,t} \tag{1}$$

The $CRD_{gi,T}$ of green index gi over the time horizon T is given by:

$$CRD_{gi,T} = \sum_{t=1}^{T} rd_{gi,t} \tag{2}$$

The average CRD, $ACRD_t$, for a green index gi across four regions k at time t is given by:

$$ACRD_{gi,t} = \frac{1}{4}\sum_{k=1}^{4} CRD_{gi,t,k} \tag{3}$$

In addition, we calculate the net CRD and net ACRD by deducting one twelfth of the corresponding ETFs' average annual costs[13] from the monthly returns of an index in our sample.

In robustness analyses, we use two other performance measures namely the Wealth Relative (WR) and Tobin's Q. The WR is the ratio of the compounded returns between a green index gi and its parent index pi at time t, estimated as follow:

$$WR_{gi,t} = \frac{\prod_{t=1}^{T}\left(1 + r_{gi,t}\right)}{\prod_{t=1}^{T}\left(1 + r_{pi,t}\right)} \tag{4}$$

A WR value greater than one indicates that an investor would be better off investing in the green index gi instead of its standard index pi over the time horizon T.

We estimate an index's Tobin's Q as the weighted sum of firm j's weight ω in index i and firm j's Tobin's Q at time t, where J is the total number of firms in the index:[14]

[13] For this purpose, we calculate the average total expense ratios of the largest ETFs which replicate green and standard indices. The average annual total expense ratios we have calculated for the standard (green) index for the World is 13 (21), for the USA 11 (20), for Europe 15 (19) and for EM 17 (25) basis points respectively.

[14] Although Tobin's Qs at the company level are annual values, the Tobin's Qs at the index level are computed monthly as the weights of constituents in an index change monthly and the composition of an index also changes over the course of a year.

$$Tobin's\ Q_{i,t} = \sum_{j=1}^{J} \omega_{i,j,t} * Tobin's\ Q_{j,t} \tag{5}$$

In a similar approach, we also estimate the Sharpe ratio and Treynor ratio of each index.

4.2.2 Returns co-movement

We estimate the Fama-French (FF) five-factor model to examine the co-movements of a green index's returns relative to the market index's returns in each region. In this analysis, we focus on the coefficient estimate of the market factor (β_1). If $\beta_1$is close to one, it indicates that the green index *i*'s returns move closely with the market index's returns. Our baseline FF five-factor model is specified as follows:

$$r_{i,t} - r_{f,t} = \alpha\ + \beta_1(r_{m,t} - r_{f,t}) + \beta_2 SMB_t + \beta_3 HML_t + \beta_4 RMW_t + \beta_5 CMA_t\ + \epsilon_t \tag{6}$$

where $r_{i,t}$ represents an index *i*'s return at time *t,* and $r_{f,t}$ is the risk-free rate at time *t*.

The independent variables include the market excess return ($r_{m,t}$ - $r_{f,t}$), Small Minus Big (SMB_t), High Minus Low (HML_t), Robust Minus Weak (RMW_t), and Conservative Minus Aggressive (CMA_t) factors at time *t*.

In addition to the baseline FF five-factor model (Fama and French, 2015), we also estimate the FF three-factor model (Fama and French, 1993), the Carhart Fama-French four-factor model (Carhart 1997), and the Carhart Fama-French six-factor model (Fama and French, 2018) for robustness tests.

To examine the co-movements between the returns of a green index and its respective standard index (in the same region), we apply the GARCH (1,1) model with a Gaussian distribution to our daily returns. This enables us to understand the correlation dynamics over time between green indices and their parent index in each region.

4.2.3 OLS index-level regressions

We carry out ordinary least squares (OLS) regressions at the index level to analyze how the effectiveness of green criteria, denoted as GCE, affects an index's financial performance. For each performance metric outlined earlier (CRD, WR, Tobin's Q), we estimate the following regression model:

$$Y_{i,t} = \alpha + \sum_{k=1}^{4} \beta_i \times GCE_{Index_i} + \sum_{g=1}^{4} \gamma_g \times HDI_{Group_{g,t-12}} + \sum_{s=1}^{10} \delta_s \times Sector_{s,t-1} \quad (7)$$

$$+\psi_1 Post\_Launch_{i,t-1} + \psi_2 logMV_{i,t-1} + \psi_3 Covid_{t-1} + \psi_4 Ukraine_{t-1} + \epsilon_t$$

where

$Y_{i,t}$: A performance measure (CRD, WR, Tobin's Q) for index i at time t, as discussed above.

GCE_Index_i : A dummy variable that takes the value of 1 if index *i* was constructed according to the criteria of a certain index. This means that GCE_PAB equals to 1 only for the PAB index while it is zero for all others, including the standard, CC, SRI, and SRI PAB indices. This logic applies to all other green index types accordingly.

$HDI\_Group_{g,t-12}$: Percentage of firms in a Human Development Indicator (HDI) group g at time $(t-12)$.[15]

$Sector_{s,t-1}$: Percentage of firms in sector s at time $(t-1)$.[16]

$Post\_Launch_{i,t-1}$: Dummy variable indicating if index i had already been launched prior to or was launched at time $(t-1)$.

lMV_{t-1}: Logarithm of the average weighted market value of index i at time $(t-1)$.

$Covid_{t-1}$: One-month lagged dummy variable for the COVID-19 period.

$Ukraine_{t-1}$: One-month lagged dummy variable for the Ukraine war period.

ϵ_t: Error term at time *t*.

Unlike ElBannan (2024), which uses a single dummy variable to represent various ESG funds, we analyze four green index types separately. Instead of grouping them in one category, we create four distinct dummy variables (GCE_Index_i). In equation (7), the standard index acts as the baseline category and is captured by the intercept (α). If the coefficient for GCE_Index_i is positive ($\beta_i > 0$) and statistically significant, it suggests that MSCI's criteria to create this specific green index type has a positive impact on its performance measure $Y_{i,t}$, even after controlling for relevant index-, sector-, country- and time-specific variables.

Our dataset includes 50 countries. Using multiple country-specific variables (for example, country dummy variables) leads to multicollinearity and would obscure the effects of our main

[15] Each year, constituents' countries in our sample are grouped into five HDI categories based on their lagged numerical values: ≤0.6, 0.6–0.7, 0.7–0.8, 0.8–0.9, and ≥0.9. The highest category (≥0.9) serves as the reference group. Since firm weights within an index change monthly and index constituents also evolve over time, the HDI group values are updated each month. The HDI data is sourced from the United Nations Development Program (UNDP).

[16] The eleven sectors include Health Care, Consumer Discretionary, Materials, Industrials, Information Technology, Financials, Consumer Staples, Energy, Utilities, Communication Services, and Real Estate which we use as the reference category.

explanatory variables, green criteria effectiveness (GCE). We avoid multicollinearity by using the Human Development Indicator (HDI) as a proxy for country-specific effects, and sector composition, which captures broad cross-sectional differences.[17]

Our primary financial performance measure is the cumulative return differential (CRD), which reflects the long-term perspective of investors. We use OLS regression as it offers clear and easily interpretable coefficient estimates which allow us to directly assess the influence of the GCEs on an index's relative return. Furthermore, we observe no significant differences in results between cumulative (CRD) and non-cumulative return measure (WR). The usage of cumulative values does not affect the main insights or conclusions of the study, suggesting that the OLS approach is suitable for our analysis.

To further validate our model, we conduct Residual versus Fitted Plots for our baseline regression of CRD. The residuals appear reasonably centered around zero across the range of fitted values, indicating no severe violations of linearity. We also examine added-variable plots, and the partial regression lines show a linear pattern for our GCE dummies. Our model further includes controls for sector composition, country characteristics (HDI), index's market capitalization, and major events (COVID-19, Ukraine war).

Most importantly, we use Newey–West standard errors in all analyses to address potential autocorrelation and heteroskedasticity.[18] This ensures that any omitted variable bias or violations of classical assumptions are minimized. We acknowledge that no model is perfect; however, the absence of strong evidence of non-linearity or severe specification errors supports the validity of our OLS analysis results.

In addition to our baseline OLS analysis of cumulative return differential, we conduct several robustness tests using alternative financial performance metrics (WR, Tobin's Q). We further analyze subsamples, including specific sub-indices, sub-regions, periods of heightened market uncertainty, and samples with two-month/three-month lags. The results remain consistent across these tests, reinforcing the reliability of our main findings.

[17] Subsequent empirical analyses show that the green indices have betas close to 1 and they all maintain high correlations (generally above 0.95) with their respective standard indices. That is, the green indices have similar systematic risk as the market index, and macroeconomic factors, such as interest rates or inflation, likely affect both green and standard indices similarly. Thus, we do not employ country-specific macroeconomic controls.
[18] Additionally, we compute White standard errors for all models and obtain consistent results.

5 Empirical results

5.1 Statistics

Table 1 presents the statistics of returns for the 20 MSCI indices in four regions over the period January 2015–December 2023 (columns 1–5), and two sub-periods of different market conditions: the pre-pandemic (market stability) period January 2015–January 2020 (columns 6–10), and the pandemic period February 2020–January 2022 (columns 11–15).[19]

Across all periods, the green indices generally demonstrate better performance than the standard index, as evidenced by greater or quite comparable mean returns. On average, the returns are highest in the USA and lowest in EM

Panel A of Figure 1 depicts the average cumulative return differentials (ACRD) for All Regions, and Panels B-E respectively show the cumulative return differentials (CRD) for the World, the USA, Europe, and EM.

As investors interested in sustainability are typically long-term oriented, our discussion below focuses on the relative performance over the medium- and long-term horizon. For All Regions, the World, and the USA, all the green indices exhibit positive CRDs (Panel A, B, and C). The ACRD (Panel A) highlights that the two indices with the most restrictive filters (SRI and SRI PAB) gain momentum and outperform their less green counterparts in the latter years of the study period. For Europe, three green indices (except CC) achieve positive CRDs throughout the entire period (Panel D). However, for EM, such outperformance was not evident, with only the SRI exceeding other indices by a large margin from year five (2020) onward.

We report the ACRD for All Regions and the CRDs for each region at a five-year horizon and over the entire study period in columns 1–2, Table 2. Furthermore, we report net ACRDs and net CRDs in columns 3–4, Table 2. Our analysis focuses on the medium to long-term horizons which suit investors who care about the long-term impacts of climate changes.

[19] We choose February 2020 as the starting point of the pandemic, as the World Health Organization (WHO) declared it a Public Health Emergency of International Concern on January 30 (WHO 2020). The period ends in January 2022, aligning with the onset of the Ukraine War in February 2022. We observe that by March/ April 2022, the strict measures previously enacted in response to the Covid-19 outbreak had started to ease, returning to pre-March 2020 levels, although this varied by country.

Table 1. Descriptive statistics of index returns, January 2015–December 2023

Panel A: World	Total: 01/15 – 12/23					Pre-Covid-19: 01/15 – 01/20					Covid-19: 02/20 – 01/22				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
	Mean	Median	SD	Min	Max	Mean	Median	SD	Min	Max	Mean	Median	SD	Min	Max
Standard	0.82*	1.4**	4.49	-13.2	12.8	0.73*	1.15**	3.35	-7.6	7.9	1.41	2.53	5.80	-13.2	12.8
Climate Change	0.91**	1.34**	4.62	-12.3	12.8	0.81*	1.09**	3.36	-7.6	7.6	1.53	2.49	5.86	-12.3	12.8
PAB	0.88**	1.54**	4.55	-12.7	12.7	0.85**	1.15**	3.30	-7.4	7.7	1.4	2.47	5.87	-12.7	12.7
SRI	0.91**	1.3**	4.52	-10.8	11.9	0.81*	1.14**	3.31	-7.6	7.7	1.59	2.78	5.70	-10.8	11.9
SRI PAB	0.9**	1.12**	4.54	-11.1	11.9	0.83*	1.02**	3.27	-7.7	7.7	1.57	2.6	5.66	-11.1	11.9
Panel B: U.S.	Total: 01/15 – 12/23					Pre-Covid-19: 01/15 – 01/20					Covid-19: 02/20 – 01/22				
	Mean	Median	SD	Min	Max	Mean	Median	SD	Min	Max	Mean	Median	SD	Min	Max
Standard	1**	1.4***	4.63	-12.7	13.1	0.92**	1.27***	3.45	-9.1	8.2	1.69	2.66	5.96	-12.7	13.1
Climate Change	1.11**	1.78***	4.79	-11.7	13.3	1.01**	1.54***	3.48	-8.9	8.1	1.85	2.85	6.06	-11.7	13.3
PAB	1.09**	1.59***	4.73	-11.9	13.2	1.1**	1.58***	3.42	-8.6	8.1	1.67	2.83	6.04	-11.9	13.2
SRI	1.13**	1.4***	4.73	-10.3	12.7	0.99**	1.36***	3.45	-8.6	7.8	2.03	3.38*	5.97	-10.3	12.7
SRI PAB	1.06**	1.21***	4.74	-11.4	12.3	0.93**	1.13***	3.45	-8.7	7.8	1.96	2.25	5.96	-11.4	12.3
Panel C: Europe	Total: 01/15 – 12/23					Pre-Covid-19: 01/15 – 01/20					Covid-19: 02/20 – 01/22				
	Mean	Median	SD	Min	Max	Mean	Median	SD	Min	Max	Mean	Median	SD	Min	Max
Standard	0.55	0.7	4.85	-14.6	17.0	0.43	0.59	3.67	-7.8	6.7	0.96	2.77	6.32	-14.6	17.0
Climate Change	0.55	0.44	4.92	-14.5	16.4	0.44	0.23	3.69	-8.2	6.2	0.94	2.99	6.35	-14.5	16.4
PAB	0.62	0.54	4.92	-14.6	16.1	0.52	0.38	3.64	-8.0	6.2	1.04	3.11	6.33	-14.6	16.1
SRI	0.68	0.75	4.75	-10.9	15.9	0.61	0.68	3.58	-7.8	7.2	1.04	3.27	6.09	-10.9	15.9
SRI PAB	0.71	1.02	4.81	-11.7	15.4	0.67	0.88	3.53	-7.7	6.9	1.18	3.17	6.10	-11.7	15.4
Panel D: EM	Total: 01/15 – 12/23					Pre-Covid-19: 01/15 – 01/20					Covid-19: 02/20 – 01/22				
	Mean	Median	SD	Min	Max	Mean	Median	SD	Min	Max	Mean	Median	SD	Min	Max
Standard	0.39	0.33	5.09	-15.4	14.8	0.47	0.24	4.57	-9.0	13.2	0.87	1.43	5.65	-15.4	9.2
Climate Change	0.37	0.23	5.15	-14.7	15.6	0.49	0.26	4.60	-8.9	13.1	0.85	1.26	5.58	-14.7	9.7
PAB	0.39	0.16	5.10	-15.8	13.3	0.47	0.13	4.56	-9.1	13.3	0.97	1.78	5.77	-15.8	9.7
SRI	0.54	0.21	5.44	-18.1	18.2	0.55	0.28	4.13	-8.7	13.6	1.46	2.05	6.86	-18.1	13.4
SRI PAB	0.36	0.18	5.12	-19.1	14.5	0.43	0.44	4.07	-8.6	13.2	1.01	2.03	6.75	-19.1	12.8

Panels A, B, C, D of this table respectively show the return statistics of the five indices examined for the World, the U.S., Europe, and Emerging Markets (EM) region. We report the mean return, median return, the standard deviation (SD) of returns, the minimum and maximum return for the total study period (January 2015 - December 2023), the pre-Covid-19 period (December 2015-January 2020), and the Covid-19 period (February 2020-January 2022). The five examined indices consist of the MSCI Standard (parent) Index, the MSCI Climate Change Index (CC), the MSCI Climate Paris-Aligned Index (PAB), the MSCI Socially Responsible Investment Index (SRI) and the MSCI Socially Responsible Investment filtered Paris-Aligned Benchmark Index (SRI PAB). The asterixis next to the mean and median returns respectively indicate the *p*-value for the t-test and Wilcoxon rank sum test showing whether the return statistics of an index are statistically different from zero. *** p-value ≤ 1%; ** 1% < p-value ≤ 5%; * 5% < p-value ≤ 10%.

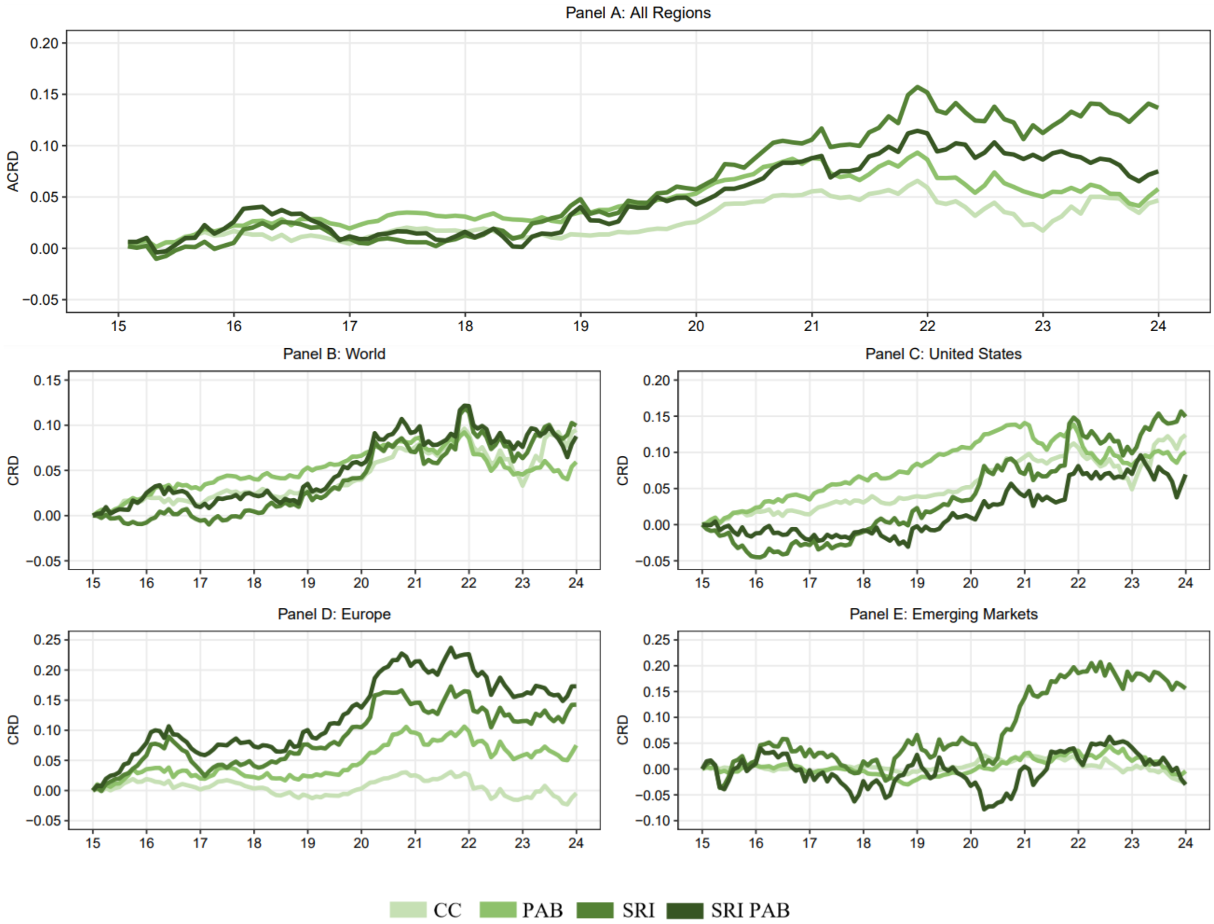


Figure 1. Cumulative Return Differential (CRD), January 2015–December 2023

Notes: This Figure presents the average cumulative return differential (ACRD) across regions and the cumulative return differential (CRD) of each MSCI green index relative to the MSCI standard/ parent index in each region over the period January 2015–December 2023. The ACRD/ CRDs do not consider costs (ETFs' expense ratios). The four green indices consist of the MSCI Climate Change Index (CC), the MSCI Climate Paris-Aligned Index (PAB), the MSCI Socially Responsible Investment Index (SRI) and the MSCI Socially Responsible Investment filtered Paris-Aligned Benchmark Index (SRI PAB).

Overall, with a few exceptions (the CC in Europe and EM, the PAB and SRI PAB in EM), the green indices in each region achieve positive net CRDs (relative to their standard index). Over our study period and across regions, the net ACRDs vary between 4% (CC) and 13% (SRI), see Panel A (column 4). The SRI is the best performer, achieving the highest net CRD in All Regions (13%, Panel A), in the World (9.2%, Panel B), in the USA (14.1%, Panel C), and in EM (14.9%, Panel E). The SRI PAB, constructed with the most restrictive criteria, delivers the best net CRD in Europe (16.9%, Panel D). Columns 5–6 of Table 2 respectively show the wealth relative (WR) and Tobin's Q of each index averaged over the entire study period.[20]

[20] As index's constituent weighting data is unavailable beyond December 2022, our analysis of index-level Tobin's Q ends in December 2022 (instead of December 2023 as for the analysis of index's returns).

Table 2. Index's financial performance measures

Panel A: Combined Regions

	ACRD (without cost)		ACRD (after deducting ETF's costs)		Wealth Relative	Tobin's Q
	5 Years (1)	Total (2)	5 Years (3)	Total (4)	Total (5)	Total (6)
Standard	-	-	-	-	1.00	2.08
CC	3.10%	4.70%	2.70%	4.00%	1.03	2.21
PAB	5.90%	5.80%	5.60%	5.10%	1.05	2.21
SRI	6.30%	13.70%	5.90%	13.00%	1.07	2.47
SRI PAB	4.70%	7.50%	4.30%	6.80%	1.05	2.41

Panel B: The World

	CRD (without cost)		CRD (after deducting ETF's costs)		Wealth Relative	Tobin's Q
	5 Years (1)	Total (2)	5 Years (3)	Total (4)	Total (5)	Total (6)
Standard	-	-	-	-	1.00	2.24
CC	4.60%	9.50%	4.20%	8.80%	1.04	2.44
PAB	7.00%	5.90%	6.50%	5.20%	1.05	2.29
SRI	4.80%	10.00%	4.40%	9.20%	1.04	2.54
SRI PAB	6.00%	8.80%	5.50%	8.10%	1.06	2.56

Panel C: The U.S.

	5 Years (1)	Total (2)	5 Years (3)	Total (4)	Total (5)	Total (6)
Standard	-	-	-	-	1.00	2.66
CC	5.90%	12.40%	5.40%	11.60%	1.06	2.86
PAB	11.30%	10.10%	10.90%	9.30%	1.08	2.91
SRI	4.30%	14.90%	3.80%	14.10%	1.04	3.00
SRI PAB	0.70%	6.90%	0.30%	6.10%	1.02	2.94

Panel D: Europe

	5 Years (1)	Total (2)	5 Years (3)	Total (4)	Total (5)	Total (6)
Standard	-	-	-	-	1.00	1.61
CC	0.80%	-0.40%	0.60%	-0.80%	1.01	1.69
PAB	5.50%	7.50%	5.30%	7.10%	1.05	1.71
SRI	11.20%	14.30%	11.00%	13.90%	1.10	2.18
SRI PAB	14.70%	17.30%	14.50%	16.90%	1.14	2.11

Panel E: Emerging Markets (EM)

	5 Years (1)	Total (2)	5 Years (3)	Total (4)	Total (5)	Total (6)
Standard	-	-	-	-	1.00	1.81
CC	1.00%	-2.80%	0.60%	-3.50%	1.00	1.86
PAB	-0.10%	-0.40%	-0.50%	-1.10%	1.00	1.93
SRI	4.80%	15.60%	4.40%	14.90%	1.08	2.14
SRI PAB	-2.60%	-3.00%	-3.00%	-3.70%	1.00	2.04

This table presents the average cumulative return differential (ACRD) in Panel A (columns 1-4), the cumulative return differential (CRD) in Panel B-E (columns 1-4), wealth relative (WR) in Panels A-E (column 5), and Tobin's Q in Panel A-E (column 6). CRD, ACRD and WR are relative to the MSCI Standard (parent) index for each region (the World, U.S., Europe, and Emerging Markets). Columns (1) and (2) of each Panel report the indices' ACRD/ CRD. Columns (3), (4) of each Panel report the net ACRD/ net CRD of corresponding synthetic Exchange Traded Funds (ETFs), which replicate our examined indices, after deducting total expense ratios. For this purpose, we calculate the average total expense ratios of the largest ETF providers for the non-sustainable and green indices, separately for each region. We then deduct 1/12 of the average annual cost from the monthly return of each index. Each performance metric (CRD, ACRD) is reported for two investment horizons starting from the study beginning in January 2015. The five-year horizon ended shortly before the spread of Covid-19 in January 2020. The analysis of index's relative returns and wealth relatives (columns 1-5) covers the period January 2015-December 2023 while the analysis of Tobin's Q (column 6) covers the period January 2015-December 2022. The calculation of the Tobin's Q of each index requires index's constituent weighting data which is only available until December 2022. The five examined indices consist of the MSCI Standard (parent) Index, the MSCI Climate Change Index (CC), the MSCI Climate Paris-Aligned Index (PAB), the MSCI Socially Responsible Investment Index (SRI) and the MSCI Socially Responsible Investment filtered Paris-Aligned Benchmark Index (SRI PAB).

Across regions, all green indices generate WRs equivalent to or greater than one, indicating that investors would be better off by investing in a green index instead of the respective standard index (column 5). Consistent with CRDs reported in columns 1–4, WRs show that the most restrictive green index, the SRI PAB, in Europe was the best performer. An investor who invested in SRI PAB Europe in January 2015 would be 14% better off over the study period than if s/he invested in the corresponding standard index. In terms of valuation, all green indices in all four regions achieve greater Tobin's Q than the respective standard index (column 6). Of particular interest, across all four regions, the two most restrictive green indices, the SRI and SRI PAB, achieve the highest Tobin's Q ranging between 2.14 and 3, and 2.04 and 2.94, respectively.

Columns (1–5) of Table 3 respectively display the Sharpe ratio (Panel A) and Treynor ratio (Panel B) of each index averaged over the study period for the World, the USA, Europe, Emerging Markets (EM), and the average ratio across regions.

Table 3. MSCI Index's Sharpe and Treynor Ratios, January 2015–December 2022

Panel A: Sharpe Ratio

	The World (1)	The U.S. (2)	Europe (3)	EM (4)	Average (5)
Standard	0.77	0.91	0.55	0.51	0.68
CC	0.81	0.94	0.59	0.53	0.73
PAB	0.8	0.96	0.61	0.63	0.75
SRI	0.84	0.97	0.68	0.60	0.79
SRI_PAB	0.82	0.91	0.69	0.46	0.78

Panel B: Treynor Ratio

	The World (1)	The U.S. (2)	Europe (3)	EM (4)	Average (5)
Standard	0.251	0.295	0.163	0.225	0.234
CC	0.265	0.306	0.176	0.238	0.246
PAB	0.257	0.308	0.182	0.241	0.247
SRI	0.270	0.308	0.221	0.254	0.263
SRI_PAB	0.270	0.291	0.212	0.216	0.247

This table presents the average Sharpe Ratio (Panel A) and Treynor Ratio (Panel B) for the regions World (Column 1), the U.S. (Column 2), Europe (Column 3), Emerging Markets (EM) (Column 4), and the average ratios across regions (Column 5). The five examined indices consist of the MSCI Standard (parent) Index, the MSCI Climate Change Index (CC), the MSCI Climate Paris-Aligned Index (PAB), the MSCI Socially Responsible Investment Index (SRI) and the MSCI Socially Responsible Investment filtered Paris-Aligned Benchmark Index (SRI PAB).

Except the SRI PAB in EM region, all green indices achieve better Sharpe and Treynor ratios than the standard index. The Sharpe ratios for the green indices range from 0.80 to 0.84 (World, column 1), 0.91–0.97 (USA, column 2), 0.59–0.69 (Europe, column 3), and 0.46–0.63 (EM, column 4), which are mostly better than the standard index's values of 0.77 (World), 0.91

(USA), 0.55 (Europe), and 0.51 (EM). On average (column 5), the green indices consistently outperform the standard index as evidenced by their greater Sharpe and Treynor ratios. The SRI is the best performer overall (with the Sharpe ratio of 0.79 and the Treynor ratio of 0.26), followed closely by the SRI PAB (0.78 and 0.247) and PAB (0.75 and 0.247).

Overall, the green indices mostly offer either comparable or greater CRD, relative to their respective standard index, over a medium- to long-term horizon. This finding holds using net CRD that considers the corresponding largest synthetic ETFs' total expense ratios which can sometimes be higher for green investment options. The evidence of the green indices outperforming the standard index is robust using alternative measures including WR, Tobin's Q, and two risk-adjusted measures namely the Sharpe ratio and the Treynor ratio.

Our findings are in line with previous literature reporting outperformance such as Fiordelisi et al. (2023) but contrast to the broader literature which does mostly not find a performance difference (Dumitrescu et al. 2023; Cunha et al. 2020; Jain et al. 2019; Kossentini et al. 2024).

5.2 Returns co-movement

A prevalent concern regarding green investment indices is the potential lack of diversification due to high exclusion rates which range between 75 and 90%.[21] To address this concern, we explore the co-movements between the returns of each green index and the market index by estimating Fama-French models. We focus on the beta coefficient estimates of the market factor of the green indices. We report the market factor's beta estimates in Table 4.

We find that the market factor's beta estimates for all green indices in all four regions are close to 1 (between 0.96 and 1.095), suggesting no substantial disadvantage in terms of volatility compared with the market index (Panel A–D, column 1). The EM has all betas' estimates exceeding 1 (ranging between 1.02 and 1.095), which indicates slightly more volatility than other regions. The SRI in EM tends to be most volatile, with beta estimates varying between 1.06 and 1.095. This is not too surprising considering its outperformance relative to its green peers and the standard index in EM region, as evidenced by a markedly higher net CRD (14.9%), a greater Sharpe ratio (0.6) and Treynor ratio (0.25).

[21] For example, the MSCI Emerging Markets SRI Filtered PAB had 177 constituents in March 2024 compared to the 1376 constituents of the standard index, representing an exclusion ratio of 89%. In the other regions, the exclusion rates for the SRI and SRI PAB are around 75%.

Table 4. Coefficient estimates of the market factor (r_m-r_f) in Fama-French Regressions, January 2015–December 2023

Panel A: The World

	Obs	FF5 (1)	FF3 (2)	C4 (3)	FF6 (4)
Standard	108	1.007***	1.004***	1.009***	1.009***
Climate Change	108	1.004***	1.019***	1.009***	1.000***
PAB	108	1.001***	1.007***	1.002***	0.998***
SRI	108	0.993***	0.993***	0.991***	0.991***
SRI PAB	108	0.995***	0.996***	1.001***	0.999***

Panel B: The U.S.

	Obs	FF5 (1)	FF3 (2)	C4 (3)	FF6 (4)
Standard	108	0.996***	0.996***	0.995***	0.995***
Climate Change	108	1.016***	1.022***	1.016***	1.013***
PAB	108	1.006***	1.011***	1.013***	1.009***
SRI	108	0.997***	0.996***	0.997***	0.998***
SRI PAB	108	0.995***	0.997***	1.005***	1.002***

Panel C: Europe

	Obs	FF5 (1)	FF3 (2)	C4 (3)	FF6 (4)
Standard	108	0.988***	0.982***	0.982***	0.986***
Climate Change	108	1.004***	0.999***	0.991***	0.995***
PAB	108	0.997***	0.997***	0.983***	0.984***
SRI	108	0.969***	0.968***	0.954***	0.956***
SRI PAB	108	0.980***	0.977***	0.973***	0.976***

Panel D: Emerging Markets (EM)

	Obs	FF5 (1)	FF3 (2)	C4 (3)	FF6 (4)
Standard	108	1.024***	1.031***	1.025***	1.017***
Climate Change	108	1.036***	1.039***	1.025***	1.021***
PAB	108	1.036***	1.034***	1.039***	1.041***
SRI	108	1.095***	1.063***	1.059***	1.092***
SRI PAB	108	1.046***	1.007***	1.000***	1.042***

This table presents the coefficient estimates of the market factor (r_m–r_f) for each index in each region, obtained from the five-factor Fama–French (FF) model (FF5, column 1, our baseline analysis), the three-factor FF model (FF3, column 2), the model with FF's three factors and Carhart's momentum factor (C4, column 3), and the model with FF's five factors and Carhart's momentum factor (FF6, column 4). The five examined indices consist of the MSCI Standard (parent) Index, the MSCI Climate Change Index (CC), the MSCI Climate Paris-Aligned Index (PAB), the MSCI Socially Responsible Investment Index (SRI) and the MSCI Socially Responsible Investment filtered Paris-Aligned Benchmark Index (SRI PAB). For brevity reason, the coefficient estimates of other factors are not presented. ***, ** and * represents p value significance at the 0.01, 0.05 and 0.10 levels respectively. The untabulated adjusted R^2s are above 93% in all models.

Next, we examine the co-movements between a green index's returns and its respective standard index's returns. We apply the GARCH (1,1) model with a Gaussian distribution to daily returns, and present the correlation dynamics for the four regions in Figure 2.

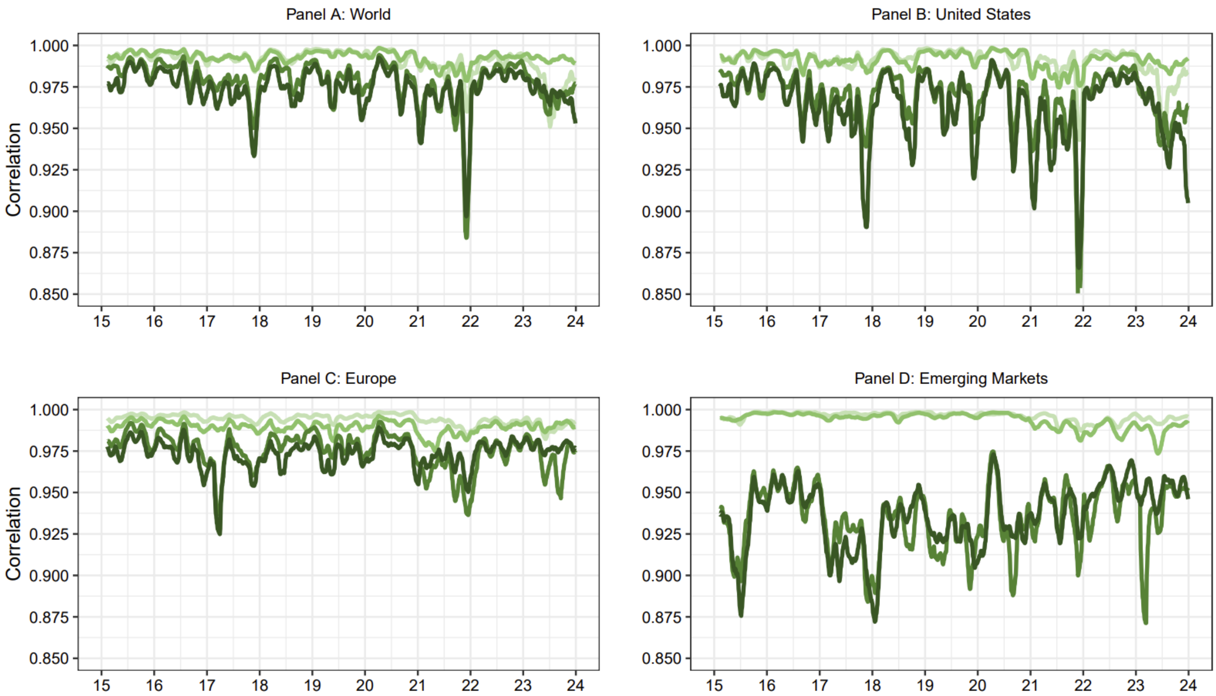


Figure 2. The dynamic conditional correlations between returns of each green index and its standard index, January 2015–December 2023

This figure features the dynamic conditional correlations between the returns of each green index and the returns of its standard parent index across four regions: World (Panel A), United States (Panel B), Europe (Panel C) and Emerging Markets (Panel B) during the period, January 2015-December 2023. The dynamic conditional correlations are obtained by estimating the GARCH (1,1) model with a Gaussian distribution using a moving average of 30 days on the obtained correlation values to smooth short-term fluctuations and reveal underlying trends.
The four green indices consist of the MSCI Climate Change Index (CC), the MSCI Climate Paris-Aligned Index (PAB), the MSCI Socially Responsible Investment Index (SRI) and the MSCI Socially Responsible Investment filtered Paris-Aligned Benchmark Index (SRI PAB).

We observe a high degree of correlations between the green indices' returns and their respective standard index's returns in all four regions, particularly for the CC and PAB indices, which often mirror the movements of their parent indices very closely at a correlation coefficient of nearly 0.98. For EU, the USA, and the World, the SRI and SRI PAB display a high correlation (around 0.97) with their corresponding parent index, suggesting a strong linkage in market behaviors. For the EM, the SRI and SRI PAB consistently maintain a correlation gap of approximately 0.05–0.1 point from their parent index. This may partly be explained by the fact that constituents in the SRI and SRI PAB in EM are, on average, of smaller size[22] and invest conservatively[23] compared with those in other regions. Within the EM, the SRI and SRI PAB have 11–12.4% constituents from South Africa, greater than other indices (5.6–6.1%). Notably,

[22] The weighted average market cap of SRI PAB (SRI) is USD 37 billion (79.9 billion) for EM and 61.3–137 billion (73.5-198 billion) for other regions (untabulated).
[23] The Conservative minus Aggressive (CMA) factor in the FF5 and FF6 regressions is statistically and significantly positive for SRI and SRI PAB indices in EM region only (untabulated).

the SRI PAB EM is more concentrated in the financial sector (34%) and less in the IT sector (13%) than other indices (23–26% in the Financial sector, and 21–25% in the IT sector).

A similar pairing is observed between SRI and SRI PAB indices. This clustering can be attributed to their compositions. These two indices are relatively similar regarding the number of excluded constituents, although the SRI PAB has an issuer cap of 5% and a more restrictive exclusion related to revenues realized from controversial activities.[24] Notably, in contrast to the CC and PAB, the SRI and SRI PAB exhibit several distinct phases where a deviation from the standard index occurred. This is not a surprise considering these indices' restrictive filters and the resulting constituents' re-weightings.

While the SRI and SRI PAB drift apart from the standard index at times, all green indices move quite closely with the standard index, as well as the market index. The finding of this analysis mitigates a concern of the green indices being undiversified, and is similar to the results of Jain et al. (2019) and Rompotis (2023).

5.3 OLS analysis: green criteria effectiveness (GCE) and index financial performance

We conduct index-level OLS analyses of CRD, WR, and Tobin's Q,[25] and report the results in columns 1–3, Table 5.[26] Each model is estimated using Newey–West standard errors. We lag all control variables by one month to mitigate concern of reverse causality. The GVIF for the group of four GCE dummy variables, estimated using DescTools in R, is less than 1.7, suggesting no serious concern of multicollinearity. Our key variables of interest are the GCE dummies, each represents the criteria MSCI uses to construct a green index type. Except the CC index, which has no significant effect on Tobin's Q (column 3), the GCE coefficient estimates for all green indices are positive and statistically significant at the 1% level. The two most restrictive indices, GCE_ SRI PAB (GCE_SRI), exert the most pronounced effects, respectively raising CRDs (column 1) by 7.7% (5.9%), WR (column 2) by 9.1% (6.4%), and Tobin's Q (column 3) by 0.635 (0.3).

---

[24] Please see a detailed discussion in "Brief overview of MSCI's methods for creating green equity indices" section and "Appendix A."

[25] Firm-level data for Tobin's Q was winsorized at the 1% and 99% level to remove outliers before being aggregated at the index level (Equation 5). Index's values of relative return measures (CRD, WR) are not affected by firm-related outliers. Untabulated additional analysis of CRD, WR, and Tobin's Q index values winsorized at the 1% and 99% level generates similarly qualitative result as the result reported in Table 5.

[26] In all regressions, the GCE values of the standard index are included in the intercept; the results do not change if the standard index observations are omitted from each regression. However, if we take out the standard index observations, one green index would be included in the intercept. Our analysis utilizes index-month observations of 20 indices over 8 years (2015–2022). Excluding 20 observations in Jan 2015 due to the use of lagged control variables, the final sample includes 1900 observations.

Table 5. OLS financial performance analysis, January 2015–December 2022

Variables	CRD (1)	Wealth Relative (2)	Tobin's Q (3)
GCE_CC	0.02***	0.037***	-0.041
	(0.007)	(0.008)	(0.035)
GCE_PAB	0.053***	0.065***	0.128***
	(0.009)	(0.01)	(0.045)
GCE_SRI	0.059***	0.064***	0.297***
	(0.004)	(0.005)	(0.019)
GCE_SRI PAB	0.077***	0.091***	0.635***
	(0.009)	(0.009)	(0.042)
Sector composite	YES	YES	YES
HDI composite	YES	YES	YES
Log of market value	YES	YES	YES
Post Launch dummy	YES	YES	YES
Ukraine dummy	YES	YES	YES
Covid dummy	YES	YES	YES
Observations	1900	1900	1900
Adj. R2	0.528	0.547	0.894

This table reports the coefficient estimates of the Green Criteria Effectiveness (GCE) of each green index obtained from the OLS index-level analysis of cumulative return differential (CRD, column 1), wealth relative (WR, column 2), and Tobin's Q (column 3). Our analysis includes all five examined indices namely the MSCI Standard (parent) Index, the MSCI Climate Change Index (CC), the MSCI Climate Paris-Aligned Index (PAB), the MSCI Socially Responsible Investment Index (SRI) and the MSCI Socially Responsible Investment filtered Paris-Aligned Benchmark Index (SRI PAB).

Each model, estimated with the Newey-West robust standard errors, includes GCE index-type dummy values, lagged firm-weighted percentages for each Human Development Indicator (HDI) group, lagged firm-weighted percentages for each sector, logarithm of lagged firms-weighted market value, lagged dummy variables that capture an index's post-launch period, the Ukraine war and the Covid-19 pandemic. Index values (CRD, WR, Tobin's Q, and control variables except dummies) are estimated using constituents' values and their weights in the index in each month.

The coefficient estimates of GCEs are reported first, followed by the standard errors (in parentheses). For brevity reason, the coefficient estimates of the intercept (the standard parent index) and control variables are not reported. ***, ** and * represents p-value significance at the 0.01, 0.05 and 0.10 levels respectively. The untabulated F-statistics for the tests of the joint significance of the coefficients of variables employed are significant at the 1% level in all model specifications.

To ensure that our results are robust, we repeat our baseline analyses using different subsamples and report the results in "Appendix B." Considering the relatively similar characteristics shared by two pairs of green indices, we first carry out separate OLS regressions of two subsamples of indices (Panel A): one with the standard, CC and PAB indices (Column 1–3), and one with the standard, SRI and SRI PAB indices (Column 4–6). Next, considering the dominance of US constituents, we analyze two subsamples of regions (Panel B): one excludes the World (Column 1–3), and one excluding the USA (Column 4–6). In panel C, we divide our sample into two sub-periods and conduct separate analyses for the market uncertainty (Column 1–3), and the market stability period (Column 4–6). Finally, to minimize concerns of reverse causality, we estimate two OLS regressions using two samples (Panel D): one with two-month lagged control variables (Column 1–3), and one with three-month lagged control variables (Column 4–6). We observe a quite consistent result that the GCEs statistically and significantly contribute to green indices' relative outperformance (CRD, WR) and, except for the GCE of CC, greater valuation (Tobin's Q). The most restrictive index, the SRI PAB, exerts the strongest impact on CRD, WR and Tobin's Q.

While the GCE of the CC index does not show a significant positive effect on Tobin's Q, this does not impact the overall implications of our study. Relative return measures CRD and WR are derived purely from market data, whereas Tobin's Q is derived based on both market and accounting data which, to some extent, is subject to earnings management. Tobin's Q shows whether an index is over-valued or under-valued (relative to the replacement cost of its assets). Tobin's Q is a valuation metric and does not capture the returns of green indices relative to their standard indices. Thus, it is not surprising that the result of the Tobin Q's analysis slightly differs from those of CRD and WR' analyses. The difference is mainly for the CC index (the least restrictive of the four examined green indices in each region). What matters most is that the significant effects of GCEs on Tobin's Q for the other three green indices indicate that market participants do recognize and reward the stricter green criteria employed—especially in the case of the SRI and SRI PAB indices, which are closely aligned with the Paris Agreement. Overall, the robust results obtained from CRD and WR analyses, and the significant effects of GCEs for SRI/SRI PAB (the two restrictive green indices that aligned with the Paris Climate Agreement) across all three measures (CRD, WR, Tobin's Q) remain the key drivers of our conclusions.

6 Conclusion

Our study analyses the financial performance 20 MSCI equity indices from 2015 to 2023 and aims at examining how the green criteria (GCE) used by MSCI to create green indices influence their financial performance. We compare the MSCI standard (parent) non-green index with four MSCI green indices in each of the four examined regions, including two that align with the Paris Agreement's objectives.

Our analyses reveal several findings. First, averaged across the four regions, the four green indices (CC, PAB, SRI, and SRI PAB) achieve better financial performance than the standard index, as demonstrated by long-term positive cumulative return differentials (CRD) and the wealth relatives (WR) being greater than 1 over the study period, particularly for the SRI indices. While we do see short-term underperformance, this is not unexpected considering the well-known trade-off between investing in sustainability and achieving immediate financial returns. Most importantly, all, except one, green indices deliver greater risk-adjusted returns (Sharpe and Treynor ratios), and higher Tobin's Qs indicating greater market valuation than the standard index. We are, however, cautious for the EM region as only the SRI perform substantially better than the standard index. Second, the returns of the four green indices are closely related to the returns of the market index and the returns of the standard index. The systematic risks of these green indices, as reflected by their beta coefficient estimates derived from Fama-French models, are similar to those of standard indices, suggesting that they do not compromise on diversification. Third, MSCI's GCEs used to construct green indices are significantly associated with their relative financial outperformance.

Our research adds valuable insights into the sustainable finance literature in several ways. First, we focus on the PAB and SRI PAB indices, making this one of the first studies to explore the financial attractiveness of Paris Agreement-Aligned Benchmark investments. Second, we analyze the financial performance of MSCI green indices compared to their standard MSCI non-green counterparts, using a rich MSCI dataset with monthly weighting data from 2015 to 2022; thereby eliminating concerns of potential effects of an index provider's characteristics. Third, we contribute to the discussion on green equity investments, showing their potential as tools for diversification and improved returns.

Our findings offer important practical insights for investors, index providers, fund managers, and policymakers. Given the observed long-term outperformance, investors seeking to align their portfolios with the Paris Agreement goals could consider ETFs tracking the MSCI PAB and SRI PAB indices (SRI PAB only for Emerging Markets). Policymakers can also use these Paris Agreement-aligned indices as reference standards for sustainable finance initiatives, for example, under the Sustainable Finance Disclosure Regulation (SFDR) Article 9 or the EU Taxonomy. They can help to verify that financial products meet defined sustainability criteria and thus increase transparency and reduce greenwashing. Using these indices as benchmarks for government pension funds or other public investment vehicles can further help to direct capital toward projects which aim to mitigate climate change. This is increasingly important considering that New Zealand Superannuation Fund and Norges Bank Investment Management, which manages Norway's sovereign wealth fund, have divested from oil companies,[27] and other sovereigns will follow suit. Our established evidence of the outperformance of the two Paris Agreement-aligned indices offers sovereign wealth funds important tools to pursue sustainable investment strategies and accelerate governments' efforts to support the transition to a low-carbon economy. Our study has certain limitations as it focuses

[27] See https://www.greenpeace.org/aotearoa/press-release/superfunds-950m-fossil-fuel-divestment-an-aha-moment-for-nz-economy/ and https://www.theguardian.com/business/2017/nov/16/oil-and-gasshares-dip-as-norways-central-bank-advises-oslo-to-divest.

on 20 MSCI indices over a specific period, which includes the COVID-19 pandemic and the ongoing war in Ukraine, potentially influencing the results due to these unusual economic conditions. The limited availability of historical constituent weighting data across regions also restricts the scope of our analysis, making it less applicable to other indices or time periods. Future research could address these limitations by examining a wider set of indices over a longer period. It could also explore how ESG/green criteria used by different index providers influence performance under different economic conditions.

References

Abate, G., I. Basile, and P. Ferrari. 2021. The level of sustainability and mutual fund performance in Europe: An empirical analysis using ESG ratings. *Corporate Social Responsibility and Environmental Management* 28 (5): 1446–1455. https://doi.org/10.1002/csr.2175.

Ang, W.R. 2015. Sustainable investment in Korea does not catch a cold when the United States sneezes. *Journal of Sustainable Finance & Investment* 5 (1–2): 16–26. https://doi.org/10.1080/20430795.2015.1042737.

Barnett, M.L., and R.M. Salomon. 2006. Beyond dichotomy: The curvilinear relationship between social responsibility and financial performance. *Strategic Management Journal* 27 (11): 1101–1122. https://doi.org/10.1002/smj.557.

Benson, C.C., N. Gupta, and R. Mateti. 2010. Does risk reduction mitigate the costs of going green? An empirical study of sustainable investing. *Southern Journal of Business and Ethics* 2: 7–25.

Bolognesi, E., A. Dreassi, M. Migliavacca, and A. Paltrinieri. 2024. Mapping sustainable investing: Exploring ambiguities and consistencies among sustainable indices. *Journal of Environmental Management* 367: 122081. https://doi.org/10.1016/j.jenvman.2024.122081.

Carhart, M.M. 1997. On persistence in mutual fund performance. *The Journal of Finance* 52 (1): 57–82. https://doi.org/10.1111/j.1540-6261.1997.tb03808.x.

Collison, D.J., G. Cobb, D.M. Power, and L.A. Stevenson. 2008. The financial performance of the FTSE4Good indices. *Corporate Social Responsibility and Environmental Management* 15 (1): 14–28. https://doi.org/10.1002/csr.144.

Consolandi, C., A. Jaiswal-Dale, E. Poggiani, and A. Vercelli. 2009. Global Standards and Ethical Stock Indexes: The Case of the Dow Jones Sustainability Stoxx index. *Journal of Business Ethics* 87 (S1): 185–197. https://doi.org/10.1007/s10551-008-9793-1.

Cortina, H., and M. Phyu. 2024. Sustainable funds continue to outgrow peers in 2023. *ISS Insights*. Retrieved from https://insights.issgovernance.com/posts/sustainable-funds-continue-to-outgrow-peers-in-2023/.

Cunha, F.A.F.D.S., E.M. De Oliveira, R.J. Orsato, M.C. Klotzle, F.L. Cyrino Oliveira, and R.G.G. Caiado. 2020. Can sustainable investments outperform traditional benchmarks? Evidence from global stock markets. *Business Strategy and the Environment* 29 (2): 682–697. https://doi.org/10.1002/bse.2397.

Dumitrescu, A., J. Järvinen, and M. Zakriya. 2023. Hidden Gem or Fool's Gold: Can passive ESG ETFs outperform the benchmarks?. *International Review of Financial Analysis* 86: 102540. https://doi.org/10.1016/j.irfa.2023.102540.

ElBannan, M.A. 2024. Returns behavior of ESG ETFs in the COVID-19 market crash: Are green funds more resilient?. *Journal of Corporate Accounting & Finance* 35 (2): 187–223. https://doi.org/10.1002/jcaf.22680.

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

Fiordelisi, F., G. Galloppo, G. Lattanzio, and V. Paimanova. 2023. Looking at socially responsible investment strategies through the lenses of the global ETF industry. *Journal of International Money and Finance* 137: 102917. https://doi.org/10.1016/j.jimonfin.2023.102917.

Folger-Laronde, Z., S. Pashang, L. Feor, and A. ElAlfy. 2022. ESG ratings and financial performance of exchange-traded funds during the COVID-19 pandemic. *Journal of Sustainable Finance & Investment* 12 (2): 490–496. https://doi.org/10.1080/20430795.2020.1782814.

French, K.R. 2024. Data Library. Retrieved from https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

Giese, G., L.-E. Lee, D. Melas, Z. Nagy, and L. Nishikawa. 2019. Foundations of ESG investing: How ESG affects equity valuation, risk, and performance. *The Journal of Portfolio Management* 45 (5): 69–83. https://doi.org/10.3905/jpm.2019.45.5.069.

Gregory, R.P., J.G. Stead, and E. Stead. 2021. The global pricing of environmental, social, and governance (ESG) criteria. *Journal of Sustainable Finance & Investment* 11 (4): 310–329. https://doi.org/10.1080/20430795.2020.1731786.

Hong, H., and M. Kacperczyk. 2009. The price of sin: The effects of social norms on markets. *Journal of Financial Economics* 93 (1): 15–36. https://doi.org/10.1016/j.jfineco.2008.09.001.

Huang, Y. 2024. Do ESG ETFs provide downside risk protection during Covid-19? Evidence from forecast combination models. *International Review of Financial Analysis* 94: 103320. https://doi.org/10.1016/j.irfa.2024.103320.

Jacob, A., and M. Wilkens. 2021. What drives sustainable indices? A framework for analyzing the sustainable index landscape. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3874437.

Jain, M., G.D. Sharma, and M. Srivastava. 2019. Can sustainable investment yield better financial returns: A comparative study of ESG indices and MSCI indices. *Risks* 7 (1): 15. https://doi.org/10.3390/risks7010015.

Jain, N., V. Mehrotra, and P. Mendiratta. 2023. Performance analysis of ESG indices: A step towards fulfilling social responsibility. *Bimaquest* 23 (1): 22–37.

Jonwall, R., S. Gupta, and S. Pahuja. 2024. Do socially responsible indices outperform conventional indices? Evidence from before and after the onset of Covid-19. *Corporate Social Responsibility and Environmental Management* 31 (5): 4995–5011. https://doi.org/10.1002/csr.2839.

Kossentini, H., O. Belhassine, and A. Zenaidi. 2024. ESG index performance: European evidence. *Journal of Asset Management* 25 (7): 653–665. https://doi.org/10.1057/s41260-024-00361-4.

Lean, H.H., and D.K. Nguyen. 2014. Policy uncertainty and performance characteristics of sustainable investments across regions around the global financial crisis. *Applied Financial Economics* 24 (21): 1367–1373. https://doi.org/10.1080/09603107.2014.925063.

Lin, X., and R.B. Swain. 2024. Performance of negatively screened sustainable investments during crisis. *International Review of Economics & Finance* 93: 1226–1247. https://doi.org/10.1016/j.iref.2024.04.001.

Managi, S., T. Okimoto, and A. Matsuda. 2012. Do socially responsible investment indexes outperform conventional indexes?. *Applied Financial Economics* 22 (18): 1511–1527. https://doi.org/10.1080/09603107.2012.665593.

MSCI. 2024a. MSCI SRI indexes methodology. Retrieved from: https://www.msci.com/documents/10199/2f754e5a-310a-0b5f-5273-5921648b1a7c.

MSCI. 2024b. MSCI SRI filtered PAB indexes methodology. Retrieved from: https://www.msci.com/documents/10199/d21129da-4fe6-9c55-81d0-eb42d4931b51.

Ng, A.C., and Z. Rezaee. 2015. Business sustainability performance and cost of equity capital. *Journal of Corporate Finance* 34: 128–149. https://doi.org/10.1016/j.jcorpfin.2015.08.003.

Omura, A., E. Roca, and M. Nakai. 2021. Does responsible investing pay during economic downturns: Evidence from the COVID-19 pandemic. *Finance Research Letters* 42: 101914. https://doi.org/10.1016/j.frl.2020.101914.

Ortas, E., R.L. Burritt, and J.M. Moneva. 2013. Socially responsible investment and cleaner production in the Asia Pacific: Does it pay to be good?. *Journal of Cleaner Production* 52: 272–280. https://doi.org/10.1016/j.jclepro.2013.02.024.

Ortas, E., J.M. Moneva, R. Burritt, and J. Tingey-Holyoak. 2014. Does sustainability investment provide adaptive resilience to ethical investors? Evidence from Spain. *Journal of Business Ethics* 124 (2): 297–309. https://doi.org/10.1007/s10551-013-1873-1.

Ortas, E., J.M. Moneva, and M. Salvador. 2012. Does socially responsible investment equity indexes in emerging markets pay off? Evidence from Brazil. *Emerging Markets Review* 13 (4): 581–597. https://doi.org/10.1016/j.ememar.2012.09.004.

Papathanasiou, S., and D. Koutsokostas. 2024. Sustainability ratings and fund performance: New evidence from European ESG equity mutual funds. *Finance Research Letters* 62: 105095. https://doi.org/10.1016/j.frl.2024.105095.

Pavlova, I., and M.E. De Boyrie. 2022. ESG ETFs and the COVID-19 stock market crash of 2020: Did clean funds fare better?. *Finance Research Letters* 44: 102051. https://doi.org/10.1016/j.frl.2021.102051.

Renneboog, L., J. Ter Horst, and C. Zhang. 2008. The price of ethics and stakeholder governance: The performance of socially responsible mutual funds. *Journal of Corporate Finance* 14 (3): 302–322. https://doi.org/10.1016/j.jcorpfin.2008.03.009.

Rompotis, G.G. 2022a. The ESG ETFs in the UK. *Journal of Asset Management* 23 (2): 114–129. https://doi.org/10.1057/s41260-021-00251-z.

Rompotis, G.G. 2022b. Environmental, social and governance responsibility, financial performance and assets: A study of exchange traded funds. *Research Papers in Economics and Finance* 6 (2): 23–49. https://doi.org/10.18559/ref.2022.2.2.

Rompotis, G.G. 2023. The performance of ESG ETFs in the US. *Capital Markets Review* 31 (2): 89–101.

Rompotis, G.G. 2024. The performance of anti-ESG ETFs in the United States. *Journal of Asset Management* 25 (7): 700–713. https://doi.org/10.1057/s41260-024-00381-0.

Schröder, M. 2007. Is there a difference? The performance characteristics of SRI equity indices. *Journal of Business Finance & Accounting* 34 (1–2): 331–348. https://doi.org/10.1111/j.1468-5957.2006.00647.x.

Statman, M. 2005. Socially responsible indexes: Composition and performance. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.705344.

Unruh, G., D. Kiron, N. Kruschwitz, M. Reeves, H. Rubel, and Meyer zum Felde, A. 2016. Investing for a Sustainable Future. *MIT Sloan Management Review*.

World Health Organization. 2020. Archived: WHO Timeline - COVID-19. Retrieved from: https://www.who.int/news/item/27-04-2020-who-timeline---covid-19.

Appendix A: Comparison of exclusion thresholds in green MSCI's indices

The MSCI SRI and SRI Filtered Paris-Aligned Benchmark (PAB) indices are designed to reflect socially responsible investing principles while aligning with sustainability and climate-focused goals. These methods impose strict thresholds on revenues derived from controversial or sensitive activities to determine company (constituents) eligibility. The thresholds vary between the SRI Index and the stricter SRI Filtered PAB Index, ensuring that investments meet higher environmental, social, and governance (ESG) standards. The following summary outlines the key thresholds for these activities based on the MSCI SRI Indexes Methodology (MSCI 2024a) and the MSCI SRI Filtered PAB Indexes Methodology (MSCI 2024b). For Controversial Weapons, companies are excluded entirely from SRI and SRI filtered PAB indices, with a 0% threshold for all revenues. For Conventional Weapons, the SRI index allows up to 5% of revenues from production and 15% from components, while the SRI filtered PAB reduces this threshold to 5% for both production and components. Civilian Firearms are also restricted, with production revenues capped at 0% and distribution revenues capped at 5% for both the SRI and SRI filtered PAB indices. Similarly, for Nuclear activities, revenues from weapons production are restricted to 0%, power generation is capped at 5%, and nuclear suppliers are limited to 15% under the SRI index. However, the SRI filtered PAB reduces the supplier threshold to 5%. Thermal Coal is excluded entirely from the SRI and the SRI PAB index while both allow up to 5% of revenues from power generation. For Tobacco, production is not allowed (0% threshold) in both SRI indices, while distribution is capped at 5%. Alcohol-related activities have more lenient thresholds. Production revenues are capped at 5% for both SRI indices, and distribution revenues are capped at 15%. Similarly, Gambling is limited to 5% for ownership revenues and 15% for services in both SRI and SRI filtered PAB indices. For Adult Entertainment, production revenues are limited to 5%, and distribution is capped at 15% under both SRI indices. Genetically modified organisms (GMOs) are restricted to a 5% threshold for production revenues. Oil and Gas activities face stricter limitations. Revenues from conventional and unconventional oil and gas activities are capped at 0% for both SRI indices. However, power generation from oil and gas is allowed up to 30% under the SRI filtered PAB index.

Appendix B. OLS robustness analysis utilizing various subsamples

Panel A: Subsamples of Indices

Variables	Sample: Standard, CC & PAB indices			Sample: Standard, SRI & SRI PAB indices		
	(1)	(2)	(3)	(4)	(5)	(6)
	CRD	WR	Tobin's Q	CRD	WR	Tobin's Q
GCE_CC	0.082***	0.114***	0.23***	NA	NA	NA
	(0.005)	(0.005)	(0.053)	NA	NA	NA
GCE_PAB	0.11***	0.136***	0.322***	NA	NA	NA
	(0.007)	(0.006)	(0.059)	NA	NA	NA
GCE_SRI	NA	NA	NA	0.074***	0.08***	0.37***
	NA	NA	NA	(0.004)	(0.005)	(0.024)
GCE_SRI PAB	NA	NA	NA	0.123***	0.137***	0.618***
	NA	NA	NA	(0.007)	(0.008)	(0.051)
Sector composite	YES	YES	YES	YES	YES	YES
HDI composite	YES	YES	YES	YES	YES	YES
Log of market value	YES	YES	YES	YES	YES	YES
Post Launch dummy	YES	YES	YES	YES	YES	YES
Ukraine dummy	YES	YES	YES	YES	YES	YES
Covid dummy	YES	YES	YES	YES	YES	YES
Observations	1140	1140	1140	1140	1140	1140
Adj. R2	0.733	0.83	0.922	0.66	0.664	0.895

Panel B: Subsamples of Regions

Variables	Sample excluding "World" Region			Sample excluding "U.S." Region		
	(1)	(2)	(3)	(4)	(5)	(6)
	CRD	WR	Tobin's Q	CRD	WR	Tobin's Q
GCE_CC	0.036***	0.038***	-0.083**	-0.013**	0.013*	-0.084**
	(0.007)	(0.008)	(0.038)	(0.006)	(0.007)	(0.037)
GCE_PAB	0.063***	0.066***	0.115**	0.012	0.033***	0.017
	(0.009)	(0.01)	(0.048)	(0.008)	(0.009)	(0.046)
GCE_SRI	0.071***	0.077***	0.332***	0.068***	0.068***	0.336***
	(0.005)	(0.006)	(0.021)	(0.004)	(0.005)	(0.021)
GCE_SRI PAB	0.094***	0.104***	0.645***	0.076***	0.092***	0.526***
	(0.009)	(0.01)	(0.043)	(0.008)	(0.009)	(0.043)
Sector composite	YES	YES	YES	YES	YES	YES
HDI composite	YES	YES	YES	YES	YES	YES
Log of market value	YES	YES	YES	YES	YES	YES
Post Launch dummy	YES	YES	YES	YES	YES	YES
Ukraine dummy	YES	YES	YES	YES	YES	YES
Covid dummy	YES	YES	YES	YES	YES	YES
Observations	1425	1425	1425	1425	1425	1425
Adj. R2	0.56	0.564	0.901	0.603	0.598	0.888

This Appendix reports the coefficient estimates of the Green Criteria Effectiveness (GCE) of each green index (the MSCI Climate Change Index-CC, the MSCI Climate Paris-Aligned Index-PAB, the MSCI Socially Responsible Investment Index-SRI, and the MSCI Socially Responsible Investment filtered Paris-Aligned Benchmark Index-SRI PAB) obtained from the OLS index-level analysis of the cumulative return differential (CRD, columns 1 and 4), wealth relative (WR, columns 2 and 5) and Tobin's Q (columns 3 and 6) for each subsample in Panels A-D. From the entire sample covering the period January 2015-December 2022, we create 8 different subsamples. *(Caption continued on next page)*

Appendix B. OLS robustness analysis utilizing various subsamples (continued)

Panel C: Subsamples of time periods

Variables	Sample market uncertainty			Sample market stability		
	(1) CRD	(2) WR	(3) Tobin's Q	(4) CRD	(5) WR	(6) Tobin's Q
GCE_CC	0.054***	0.009*	0.002	0.008	0.017	-0.219***
	(0.006)	(0.005)	(0.073)	(0.01)	(0.011)	(0.048)
GCE_PAB	0.136***	0.034***	0.146	0.038***	0.044***	-0.069
	(0.012)	(0.009)	(0.125)	(0.01)	(0.011)	(0.046)
GCE_SRI	0.112***	0.04***	0.377***	0.025***	0.026***	0.292***
	(0.007)	(0.005)	(0.070)	(0.004)	(0.004)	(0.02)
GCE_SRI PAB	0.152***	0.043***	0.515***	0.041***	0.047***	0.395***
	(0.008)	(0.006)	(0.081)	(0.012)	(0.014)	(0.057)
Sector composite	YES	YES	YES	YES	YES	YES
HDI composite	YES	YES	YES	YES	YES	YES
Log of market value	YES	YES	YES	YES	YES	YES
Post Launch dummy	YES	YES	YES	YES	YES	YES
Observations	700	700	700	1200	1200	1200
Adj. R2	0.775	0.487	0.857	0.481	0.474	0.847

Panel D: Sample with two- and three-month lagged control variables

Variables	Sample with two-month lagged control variables			Sample with three-month lagged control variables		
	(1) CRD	(2) WR	(3) Tobin's Q	(4) CRD	(5) WR	(6) Tobin's Q
GCE_CC	0.024***	0.042***	-0.034	0.027***	0.046***	-0.058
	(0.007)	(0.008)	(0.035)	(0.007)	(0.008)	(0.036)
GCE_PAB	0.06***	0.073***	0.14***	0.064***	0.077***	0.105**
	(0.009)	(0.01)	(0.047)	(0.008)	(0.011)	(0.048)
GCE_SRI	0.063***	0.068***	0.331***	0.065***	0.071***	0.352***
	(0.005)	(0.005)	(0.023)	(0.005)	(0.005)	(0.024)
GCE_SRI PAB	0.081***	0.095***	0.653***	0.085***	0.099***	0.629***
	(0.009)	(0.01)	(0.045)	(0.008)	(0.01)	(0.045)
Sector composite	YES	YES	YES	YES	YES	YES
HDI composite	YES	YES	YES	YES	YES	YES
Log of market value	YES	YES	YES	YES	YES	YES
Post Launch dummy	YES	YES	YES	YES	YES	YES
Ukraine dummy	YES	YES	YES	YES	YES	YES
Covid dummy	YES	YES	YES	YES	YES	YES
Observations	1880	1880	1880	1860	1860	1860
Adj. R2	0.524	0.542	0.896	0.522	0.54	0.892

(Caption continued from previous page) In Panel A, we examine two subsamples of which one consists of the standard and two less strict CC and PAB indices (columns 1-3), and the other including the standard and two more strict SRI and SRI PAB indices (columns 4-6). In Panel B, we analyse two subsamples of which one excludes the "World" region (columns 1-3), and one exclude the "U.S." region (columns 4-6). In Panel C, we look at two different sub-periods: market uncertainty (columns 1-3) and market stability (columns 4-6). The market uncertainty period covers the Ukraine war and the Covid pandemic while the market stability period covers the remaining months of our study period. In Panel D, we explore a sample with two-month lagged control variables (columns 1-3) and a sample with three-month lagged control variables (columns 4-6). In both samples, the index-level HDI composite is lagged by a year. We estimate all models with the Newey-West robust standard error. The coefficient estimates of GCEs are reported first, followed by the standard errors (in parentheses). For brevity reason, the coefficient estimates of the intercept and control variables are not reported. ***, ** and * represents *p*-value significance at the 0.01, 0.05 and 0.10 levels respectively.

Paper 2 – Investing in a Low-Carbon Transition: Carbon Footprint Savings with Green MSCI Indices

Abstract*

Green equity index providers claim that they can play a central role in allocating capital towards decarbonization. This study verifies this claim by analysing the environmental performance of 20 MSCI equity indices comprising 5267 constituents in 50 countries over the period 2015–2022. In each region (the World, the U.S., Europe, Emerging Markets), we analyse four MSCI green indices, namely Climate Change (CC), Paris-Aligned Benchmark (PAB), Socially Responsible Investment (SRI), and SRI Filtered PAB (SRI PAB), relative to the MSCI standard index. Our aggregated index-level regression reveals that the green criteria employed by MSCI statistically and significantly contributed to CO_2 emissions reductions for all examined green indices, and better environmental/ ESG ratings for the CC, SRI, and SRI PAB indices. This relationship is robust to aggregated index-level control variables, various subsamples, and alternative emissions measures provided by Refinitiv, Bloomberg, and CDP. An investment of US$1000 in January 2015 in the CC, PAB, SRI, and SRI PAB indices (instead of the standard index) respectively resulted in an average reduction of 982 kg (−53%), 1429 kg (−77%), 1147 kg (−61%), and 1355 kg (−73%) CO_2 emissions at the end of December 2022. The robust environmental outperformance of Paris Agreement–aligned indices suggests that they should be embedded in benchmark regulations and granted preferential capital treatment. Institutions should set low-carbon targets to guide asset managers towards green ETF investments to accelerate the low-carbon economy transition. Our study also stresses the importance of improving transparency, standardizing ESG performance metrics, and rating criteria to support green investing.

Keywords: MSCI; Green index; Carbon emissions; ESG scores; Environmental rating; Controversy rating

* This paper is joint work with Huong Dang and Thomas Brückner. It has been published as: Heldmann, J., H. D. Dang, and T. Brückner. 2025. “Investing in a low-carbon transition: Carbon footprint savings with green MSCI indices.” *Journal of Environmental Management* 393: 126022. https://doi.org/10.1016/j.jenvman.2025.126022. This version includes minor typographical and formatting adjustments. The content remains unchanged.

1 Introduction

Amid growing concerns about climate change, market participants have been showing great interest in investment options that would positively contribute to carbon emissions reduction and foster a low-carbon transition. Green equity Exchange-Traded Funds (ETFs)[1] have emerged as appealing vehicles for investors who seek to align their portfolios with their environmental goals and ethical values. These ETFs replicate the performance of green indices, which are constructed based on sustainability metrics such as CO_2 emissions and environmental ratings. Green equity index providers claim that they can play a central role in allocating capital towards decarbonization (Hamill and Wu, 2023; Beyhan and Kennedy, 2025; MSCI Inc., 2022). However, few studies have verified this claim.

Research on the environmental performance of green ETFs or mutual funds is limited and presents mixed results. D'Ecclesia et al. (2022) and Abouarab et al. (2022) report poor or no improvement in environmental performance compared to conventional funds, while Meyers, Ferrero-Ferrero and Muñoz-Torres (2024) and Jacob and Wilkens (2021) find that sustainable funds/indices generally outperform their standard peers. Inconsistent findings across studies likely arise from differences in samples, geographic focuses, study periods, benchmarks, and the choice of environmental performance measures. In addition, most empirical studies categorize various ESG and SRI ETFs into a single group, treating them equivalently despite their different underlying benchmarks and exclusion construction criteria. The literature is particularly scarce regarding the relationship between the construction criteria of indices aligned with the Paris Agreement and their environmental performance.

This study is motivated by a lack of solid empirical evidence on whether "green" investment labels translate into real-world carbon emissions reductions. We examine a set of green and standard MSCI indices, which have been widely used as references for green equity ETFs.[2] Recent regulatory change stresses the need to analyze broad-based market indices such as those provided by MSCI, which function as critical benchmarks.[3] We focus on MSCI equity indices that are aligned with the Paris Agreement and address two research questions: (i) How

[1] Many different terms are currently used to describe sustainable investments, for example, Environmental, Social & Governance (ESG), Socially Responsible Investment (SRI), sustainable, ethical, green. For simplicity, this study refers to these as green investments.
[2] Our analysis of the worldwide equity ETF market via Morningstar in April 2024 for Sustainable Finance Disclosure Regulation (SFDR) fund types Articles 8 and 9 shows an MSCI-provided ETF market share of 75.7 % (US$ 605 billion). When limited to the selected regions (World, U.S., Europe and Emerging Markets), the share dropped marginally to 75.4 % (US$531 billion).
[3] Since October 2022, the U.S. Securities and Exchange Commission has required fund companies to list an applicable broad-based market index for U.S. Mutual Funds and ETFs. For more details, please see https://www.sec.gov/investment/tailored-shareholder-reports-mutual-funds-etfs.

did green MSCI equity indices perform environmentally compared with their respective standard (parent) index?; and (ii) Did the green criteria employed by MSCI to construct each green equity index statistically and significantly affect its environmental performance, after considering its firm-weighted size-, sector-, and country-specific controls?

Our analyses of 20 MSCI equity indices cover four regions, namely: the World, the U.S., Europe, and Emerging Markets (EM). In each region, the MSCI standard (parent) index and four green-oriented indices, namely, Climate Change (CC), Paris-Aligned Benchmark (PAB), Socially Responsible Investment (SRI), and SRI Filtered PAB (SRI PAB) are examined. The PAB and SRI PAB indices are directly aligned with the Paris Agreement's objective; however, they have not been analyzed in the related literature. Our index dataset ranges from January 2015[4] to December 2022 and consists of 5267 unique constituents spanning 11 sectors across 50 countries, resulting in 1,076,340 firm-month observations. The rich history of MSCI indices' constituents allows us to perform a more comprehensive and robust analysis than would have been possible using comparable ETFs.

This study markedly differs from recent research on ETFs' environmental performance, such as Popescu et al. (2023). First, our access to MSCI's monthly index constituent data, including historical constituents of 20 indices across four regions over a long time period[5] allows for an in-depth index-level analysis, utilizing an index's firm-weighted data and controlling for sector- and country-specific characteristics. The monthly historical weighting data of constituents enables us to accurately compute how much carbon emissions can be reduced by investing in a green index (instead of its standard index). Second, this study examines four sets of region-specific MSCI green indices that employ mild to strict exclusion criteria. Our analysis of separate green indices (instead of a generic ESG/SRI category) provides insights into the attractiveness of each as a vehicle offering substantial carbon reductions over the long term.

This study is related to Jacob and Wilkens's (2021) unpublished study, which analyses four ESG and four carbon-focused MSCI World Indices from 2011 to 2019. Our analysis, however, extends their study by covering a more recent time period, which includes the COVID pandemic and the Ukraine war, incorporating an extensive set of 16 green MSCI equity indices, including

[4] The analysis begins in January 2015 when complete data on all indices and their individual constituents, including monthly weightings, first become available.

[5] Normally, only the most recent data are listed in MSCI's latest index factsheet. The MSCI index dataset goes back to January 2015 which is prior to the actual launch dates of 12 out of 20 examined indices; however, it is based on the same methods and principles applied to the past. No study to date has utilized such rich historical constituent data of MSCI equity indices over a period spanning both the pandemic and the Ukraine war.

those aligned with the Paris Climate Agreement, and examining three additional major regions: Europe, the U.S., and EM, alongside the global market.[6] Furthermore, this study utilizes multiple CO_2 emissions measures from three CO_2 data providers to enhance the robustness of the key findings. Most importantly, we conduct index-level regression analyses that utilize an index's constituent data and directly assess the effectiveness of criteria employed by MSCI in constructing each green index. This study enriches the limited literature on the performance of green indices, particularly those aligned with the Paris Climate Agreement.

We estimate index-level environmental metrics using an index's monthly constituent weights and constituents' corresponding metrics. Four dummy variables are employed in the OLS analyses, each capturing the effectiveness of green criteria (GCE) employed by MSCI to construct a green index type.[7] A key challenge in this study is to identify appropriate environmental performance measures, given the lack of a standardized framework. Since carbon emissions align with the objectives of the Paris Agreement, and considering data quality and availability, Scope 1 + 2 carbon emissions are selected as the primary measure. As long-term CO_2 emissions reduction is a priority for investors interested in sustainability, our baseline model uses cumulative size-adjusted carbon emissions published by Refinitiv (REF), Bloomberg (BB), and the Carbon Disclosure Project (CDP).[8] The main data provider is REF, as its data has been widely used and it has the best coverage of constituents in our sample. Additionally, alternative measures of emissions (non-cumulative size-adjusted, non-cumulative unadjusted, CO_2 intensity) provided by REF, BB, Carbon Disclosure Project, and various ESG/ environmental rating scores issued by REF and BB are employed for robustness tests. The employment of rating scores from the two providers is needed, considering the observed divergences across rating providers (Berg et al., 2022).[9]

The environmental performance analysis provides compelling evidence supporting the attractiveness of green indices. If an investor had invested US$1000 in January 2015 in the green CC, PAB, SRI, or SRI PAB indices (instead of the standard index), averaged across all regions, they would have accumulated substantial CO_2 Scope 1 + 2 emissions savings (Refinitiv data) of 982 kg (− 53 %), 1429 kg (− 77 %), 1147 kg (− 61 %), and 1355 kg (− 73 %), respectively, by December 2022. Furthermore, the GCEs employed by MSCI statistically and

[6] Jacob and Wilkens (2021) focus only on the World region. Of the 16 green MSCI indices we examine, only the World SRI Index overlaps with Jacob and Wilkens's (2021) unpublished study.
[7] Our use of a dummy variable is similar to that of Popescu et al. (2023). They use an SRI dummy to isolate the impact of various SRI ETFs in their analysis of ETFs' environmental performance.
[8] MSCI-published sustainability measures are not used in any regressions to mitigate reverse-causality concerns.
[9] According to Berg et al. (2022), differences in measurement (56 %) and scope (38 %) account for most of the divergences across providers, whereas weighting differences account for just 6 % of the variance.

significantly lead to sizeable reductions in carbon emissions for all examined green indices, and better ESG/environmental ratings for all, except PAB, indices. Our key findings are robust to additional analyses using one-, two-, and threemonth-lagged control variables, alternative emission/rating measures, and various subsamples. Our results align with those of Jacob and Wilkens (2021) but contrast with those of Popescu et al. (2023) and D'Ecclesia et al. (2022).

The findings suggest that ETFs which would replicate the examined green indices in the World, U.S., and Europe could serve as a viable strategic investment option for investors who are concerned about the potential environmental impacts of their investments. This study enriches the debates on the moral benefits of green investment and highlights the importance of integrating green criteria into the index creation process. For policymakers and regulatory bodies, this study emphasizes the importance of improving transparency, standardizing ESG disclosures and harmonizing ESG rating criteria to support the growth of green investing. Most importantly, the substantial CO_2 reductions achieved by the examined green indices demonstrate how targeted financial strategies can meaningfully advance global sustainability efforts and accelerate the transition to a low-carbon economy, particularly in alignment with the objectives of the Paris Agreement.

The rest of this paper is structured as follows. Section 2 reviews the literature and proposes research questions. Section 3 describes the data and methods. Section 4 discusses the empirical results, and Section 5 concludes with the key findings.

2 Literature review and research questions

2.1 Legitimacy and greenwashing in ESG index construction

The legitimacy theory states that organisations try to align themselves with societal norms, values, and expectations to secure acceptance and continued access to resources (Suchman, 1995). In the context of green investment, index providers face growing demands from investors who wish to align their investments with climate goals such as those outlined in the Paris Agreement. In response, providers construct and promote "green" indices that appear to align with these expectations. Similarly, financial institutions have shown a preference to label themselves or their investment products as green to signal alignment with environmental goals (Kaustia and Yu, 2021). This preference has contributed to the rise of ESG and climate-related equity indices and mutual funds in recent years (Morningstar Manager Research, 2024).

Legitimacy, however, can be achieved in both substantive and symbolic ways. Substantive legitimacy reflects actual alignment with environmental norms, for example, a reduction in CO_2

emissions compared with the parent index. Symbolic legitimacy may involve adopting green labels such as *Climate Change* or *Socially Responsible Investment* without achieving environmental improvements.[10]

The distinction between substantive and symbolic legitimacy is central to concerns about greenwashing, which refers to the practice of making unsubstantiated or misleading environmental claims (Lubloy ´ et al., 2025; Delmas and Burbano, 2011). In investment, greenwashing occurs when funds or indices are branded as sustainable without delivering corresponding environmental benefits. From a theoretical standpoint, greenwashing can be interpreted as a legitimacy-seeking behavior: index providers may use sustainability language and ESG labels as a way to enhance reputation and attract capital inflows, even if the index constituents do not exhibit superior environmental performance (Dumitrescu et al., 2022; Abouarab et al., 2022).

To ensure that green indices/funds genuinely contribute to sustainability and do not engage in greenwashing, it is necessary to assess their environmental performance. The next section will review the related literature.

2.2 Literature review

Our study focuses on equity investment, as investors who care about sustainability are typically long-term-oriented, and equities outperform bonds over a long-term horizon. Below, we briefly review several studies that examine the environmental performance of green equity funds/ indices which to date report mixed findings. Dumitrescu et al. (2022) analyse a sample of 3227 U.S. funds from 2016 to 2022 and find that 30 % of ESG funds make explicit green claims that are not supported by fund activities and are thus engaging in greenwashing. Abouarab et al. (2022) also analyse sustainable U.S. funds for a period of ten years and document that sustainable funds are not able to decrease their emissions relative to conventional funds, which Abouarab et al. (2022) consider an indication of greenwashing in these funds.

Popescu et al. (2023) find that 38 % (24 %) of SRI funds are more exposed to carbon-intensive firms than the MSCI World Index (MSCI Europe Index).[11] D'Ecclesia et al. (2022) evaluate the environmental performance of ten green ETFs based on the Environmental Score

[10] Kartal et al. (2024) use a multilayer perceptron to uncover the complex, nonlinear ties between ESG disclosures and ESG scores for 66 firms in the Borsa Istanbul Sustainability Index. They show how index creators could use algorithmic weightings to signal "green" credibility without delivering real environmental improvements.

[11] Popescu et al.'s (2023) analysis of SRI is based on the year 2019 and captures a snapshot of history. They do not employ self-reported firm-specific emissions data as it covers only 17 % of their sample. Instead, they estimate the greenhouse gas emissions by a detailed country and sectoral breakdown of companies' revenue combined with regionalized industry emissions factors.

and the carbon intensity and document poor performance. Conversely, Meyers et al. (2024) report better environmental performance in terms of sustainability ratings and risk for 92 sustainable funds.

The only study that explores equity indices, Jacob and Wilkens (2021), observes a reduction in CO_2 intensity in four ESG and four carbon-focused MSCI World Indices compared to their parent index, although the reduction varies substantially. This limited evidence underscores the need for a thorough empirical analysis of the environmental performance of green equity indices, particularly those that align with the Paris Agreement's objective. The literature is rather silent regarding the relationship between green criteria employed by index providers and indices' environmental performance, which is partly due to a lack of transparency in index construction methods.

2.3 Research questions and potential contributions

Equity indices are the building blocks of ETFs, which are popular investment options thanks to their low costs.[12] Index providers like MSCI publish a clear and thorough methodology for their equity index construction. In terms of global market share, ETFs that replicate MSCI indices account for over 70 % of all ETFs, as discussed in the introduction section. This study therefore focuses on MSCI equity indices and conducts index-level analyses to address the following two research questions.

1. Did green MSCI equity indices achieve better environmental performance than their standard indices?
2. Did the green criteria (GCE) employed by MSCI to construct each green equity index statistically and significantly affect its environmental performance, after considering its firm-weighted size-, sector-, and country-specific controls?

In addressing the above questions, this study makes use of a unique monthly index constituent weighting dataset and creates accurate index-level carbon emissions based on firms' self-reported emissions. Unlike Jacob and Wilkens (2021), this study conducts regression analyses of the GCE employed by MSCI in constructing each green index type. The time series analyses cover a relatively long period including the Ukraine war and the COVID-19 pandemic as well as a period of high inflation in major markets, which enhances the robustness of our findings.

[12] According to Morningstar data, ETFs that track green (and standard) indices or slightly modified MSCI index benchmarks have a total fund size of US $207 billion (US$1014 billion), while ETFs that track exactly our selected MSCI green (and standard) index benchmarks manage US$63 billion (US$867 billion) as of April 2024.

This study enriches the literature in four aspects. First, it examines four region-specific MSCI green equity indices (CC, PAB, SRI, SRI PAB) separately, rather than grouping them under a generic ESG/SRI category. Second, it focuses on the PAB and SRI PAB indices, which are directly aligned with the Paris Agreement's objectives, providing relevant insights for investors aiming to reduce carbon emissions. Third, it analyses MSCI green indices relative to their MSCI standard index within each region, controlling for provider-specific characteristics and offering a nuanced understanding of each green index's performance. Fourth, it provides solid empirical evidence on the effects of green index construction criteria employed by MSCI and the environmental performance of its green equity indices.

3 Method

3.1 Data

3.1.1 Constituents' data

MSCI generously provided us with a dataset including detailed monthly constituents and constituents' weights in each of the 20 indices from January 2015 to December 2022. Thus, all our regression analyses cover the period 2015–2022. The dataset includes 5267 unique constituents across 11 sectors and 50 countries in four regions. We meticulously map 5034 constituents to their International Securities Identification Numbers (ISINs), which represent 99.98 % of total index weights and ensure high coverage. Constituent-specific data are then obtained from Refinitiv Eikon's Datastream, covering nearly 100 % of our sample. This study is the first to make use of comprehensive MSCI index constituent weighting and returns data for all 20 indices including the Paris Agreement-aligned indices over a relatively long period.[13]

The green indices have varying degrees of overlap with the standard index, with the least strict CC indices covering 91.6–95.4 %, the PAB indices covering 51.4–70.5 %, and the stricter SRI and SRI PAB indices covering 16.1–26.0 % and 13.9–24.9 %, respectively, across the regions.

Appendix A describes the methodology MSCI employs to construct the four examined green index categories and provides a comprehensive discussion of their conceptual and computational distinctions.

[13] Heldmann et al. (2025) utilise the same set of green MSCI and standard indices; however, their analysis is limited to the financial performance of green indices.

3.1.2 Carbon emissions metrics

To examine the environmental performance of each index, we utilize Total CO_2 Emissions (Scope 1 + 2) as the primary measure considering its data availability and its alignment with the Paris Agreement's objectives. Firms report emissions in sustainability reports, making this metric more accessible than ESG rating scores. Our baseline analysis utilizes Refinitiv (REF) data due to its extensive coverage.[14] For robustness analysis, we employ REF's Direct CO_2 Emissions (Scope 1), REF's CO_2 Intensity, and the corresponding carbon emissions data from Bloomberg (BB) and the Carbon Disclosure Project (CDP).[15] Variables are defined as in Appendix I. Scope 3 emissions are excluded due to limited coverage (50–60 % of the sample) and reliability concerns raised in several studies (Popescu et al., 2023; Busch et al., 2022).

Firm-specific data for CO_2 emissions sometimes raise concerns due to their inaccuracy (Bajic et al., 2023). Thus, several steps are conducted to check outliers in the data and enhance the reliability of our emissions dataset. First, annual changes in emissions are calculated to identify companies with extremely high increases or decreases in emissions. This step detects and corrects a common error when companies report CO_2 emissions (in kilograms) in units of a thousand or million in their sustainability reports, but emissions are not correctly scaled in the database. Second, sector quantiles are examined to pinpoint unusually large or small firm-level values, which are then verified against the respective firms' sustainability reports. Third, monthly constituent-weighted CO_2 emissions of each index are examined. Large monthly jumps were mainly due to either the inclusion of a company from the Energy, Materials or Industrial sectors or inaccurate data. This step helps us spot constituents with high weights in an index, which led to unusually large changes in an index's emissions. Inaccurate or suspicious data are replaced manually with the data obtained from the sustainability report of the concerned firm, and if no data is available, the concerned observation is deleted.

Previous studies report that emissions values can vary across different databases (Papadopoulos, 2022). However, we find that the size-adjusted[16] firm-level Scope 1 + 2 emissions correlations across the three providers in our data are very high (0.97–0.98) and statistically significant at the 1 % level. For the size-adjusted aggregated index level, the scope

[14] For a summary of studies using Refinitiv's environmental and emission data, see de Villiers, Jia and Li (2022).
[15] Refinitiv and CDP each offer combined Scope 1 + 2 emissions intensity. For Bloomberg, Scope 1 + 2 emissions and Scope 1 + 2 emissions intensity each is estimated as the sum of the corresponding Scope 1 and Scope 2 values. For CDP, the Scope 1 + 2 emissions are the sum of Scope 1 and Scope 2 emissions. Note that there are two methods of calculating the emissions for energy consumption (Scope 2), either using the market-based approach based on specific energy suppliers' direct emissions or the location-based approach based on the average grid emissions for a geographical region. This study employs the market-based CDP Scope 2 emissions.
[16] A firm's emissions were divided by its market capitalization. This size-adjusted metric provides a normalized measure of carbon intensity relative to the company's economic value.

1 + 2 emissions correlations are 0.97 between REF and BB, 0.72 between REF and CDP, and 0.79 between CDP and BB[17]; all correlations are statistically significant at the 1 % level.

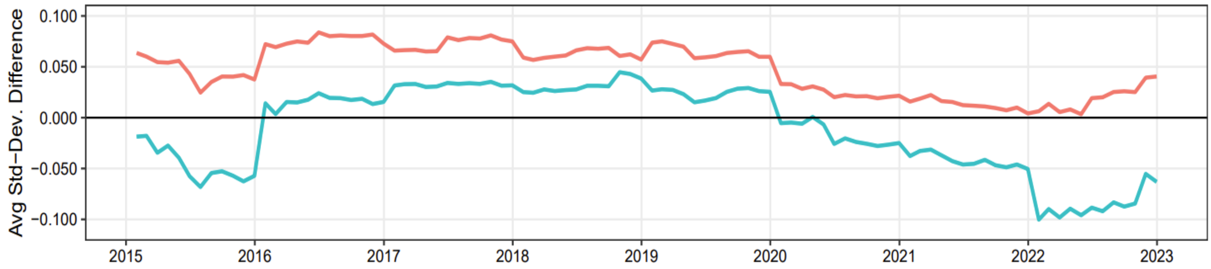
We also calculate the average monthly difference in an index's firm weighted emissions reported by the three providers. For ease of economic significance interpretation, the differences are quantified in terms of the standard deviations (std devs) of Refinitiv's emissions (our primary emissions measure), as detailed in Figure 1 (next page). On average, the differences between the three providers are quite small. For the most comprehensive metric, scope 1 + 2, the average differences across providers measured in std devs of REF emissions vary within a narrow range. Overall, we do not observe substantially large differences in an index's firm-weighted emissions reported by the three providers.

3.1.3 ESG/environmental rating metrics

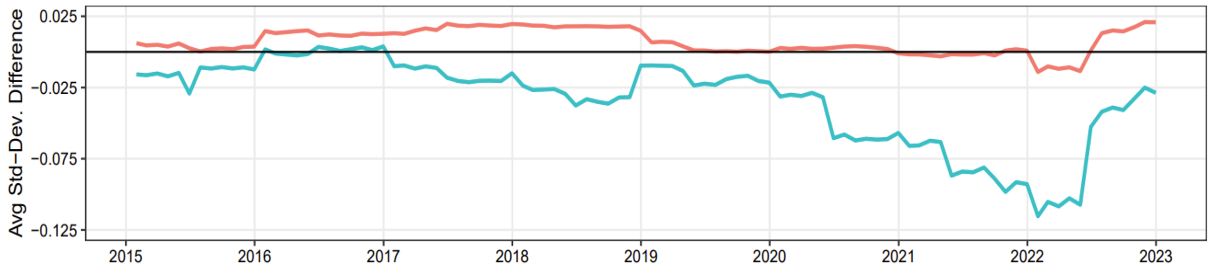
In robustness analyses, rating scores, including REF's Environmental (ENV) Score, ESG Score, and ESG Controversies Score, as defined in Appendix I, are employed. These ratings are widely utilized in sustainability studies (Flammer, 2021; Dai et al., 2021). Additionally, BB's ENV Score and ESG Score are used.[18] Similar to emissions data, rating data covers the period 2015–2022. ESG scores vary significantly across providers (Berg et al., 2022),[19] with providers like REF often adjusting scores retrospectively (Berg et al., 2021). Therefore, we cautiously use ratings as robustness tests for our CO_2 emissions results.

[17] Firm-level correlations are higher than index-level correlations as each index assigns different weights to its constituents (firms), with larger firms having more influence on the overall index performance, leading to divergences at the index level.
[18] Unlike Refinitiv, Bloomberg does not provide an ESG Controversies Score.
[19] The correlations for ENV (ESG) rating scores between REF and BB are 0.43 (0.49), REF and MSCI 0.25 (0.46), and BB and MSCI 0.23 (0.37).

Panel A: Scope 1+2



Panel B: Scope 1



Panel C: Scope 2

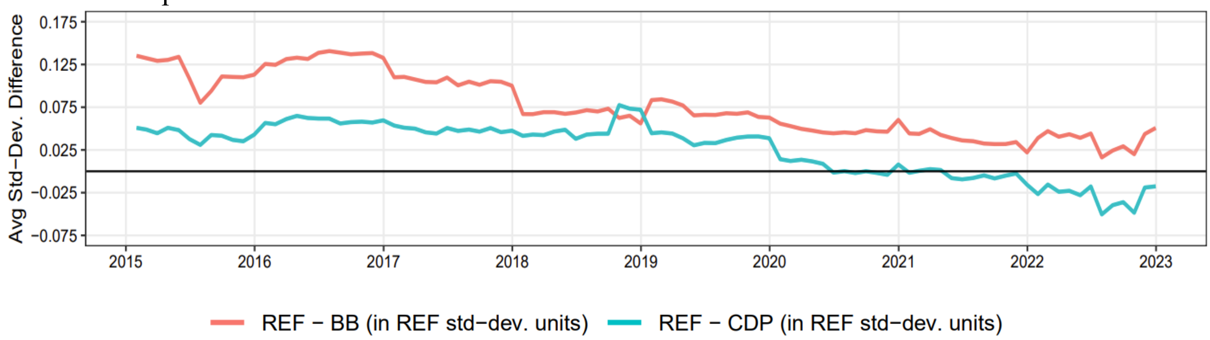


Figure 1. Average difference in index's CO_2 emissions across providers, January 2015-December 2022

This figure presents the average monthly difference in an index's size-adjusted CO_2 emissions reported by Bloomberg (BB), Carbon Disclosure Project (CDP) and Refinitiv (REF). For each month, we calculate the size-adjusted CO_2 emissions of each index in each region. We then average the index's monthly emissions across all four regions (the World, U.S., Europe, and Emerging Markets) and across all five examined indices namely the MSCI Standard Index, the MSCI Climate Change Index (CC), the MSCI Climate Paris-Aligned Index (PAB), the MSCI Socially Responsible Investment Index (SRI) and the MSCI SRI Filtered PAB Index (SRI PAB).

This analysis covers the period January 2015-December 2022, with one exception (due to the low number of matching constituents for the Emerging Markets (EM), EM observations for June 2015-December 2015 are not included). To adjust for the differences in emissions scaling, Bloomberg's emissions are scaled by a factor of 1000 and CDP's by a factor of one million so that they have the same unit (tonnes) as the unscaled emissions reported by Refinitiv. Variable definitions are provided in Appendix I.

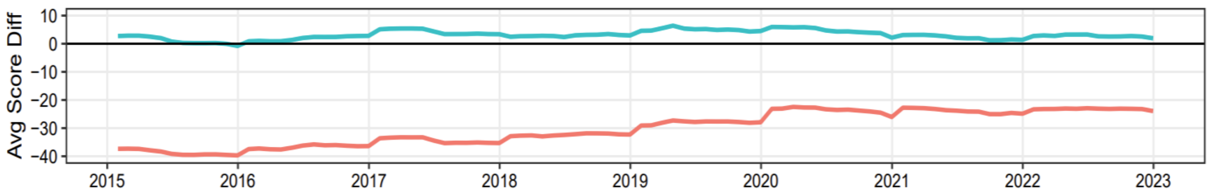
Panel A, B, and C respectively show the average differences in Scope 1+2, Scope 1, and Scope 2 emissions measured in terms of the standard deviations (std devs) of REF emissions. As REF emissions are our primary measure, we convert emissions differences into the std devs of REF emissions for ease of economic significance interpretation.

Untabulated analysis shows that the mean differences in the Scope 1+2 (Scope 1) emissions for (REF – BB) and (REF – CDP) are 0.048 (0.007) and -0.011 (-0.034) std dev of REF emissions, respectively. The average difference in Scope 1+2 (Scope 1) emissions between (REF – BB) and (REF – CDP) ranges between 0.003 and 0.084 (-0.014 and 0.021) std dev, and between -0.1 and 0.045 (-0.115 and 0.004) std devs, respectively.

To mitigate endogeneity concerns, MSCI emissions and ratings are not used in any OLS regressions. However, we still examine the divergences between MSCI and REF/BB ENV ratings to understand any potential influences on our analysis. Figure 2 illustrates ENV rating differences in raw units (Panels A, B) and in standard deviations (std dev) of MSCI ratings (Panels C, D). BB ratings are significantly lower than MSCI's; the difference (BB-MSCI) ranges between −39.7 and −22.5 rating points, whereas (REF-MSCI) ratings are mostly positive, up to 6.4 points (Panel A).[20]

Excluding firms with zero ENV scores (Panel B) minimally alters the divergence between BB and MSCI ratings.[21] Similar patterns appear when the divergences are measured in units of std devs of MSCI ENV ratings (Panels C, D, Figure 2).[22] ENV rating distributions also markedly differ: REF skews towards higher ratings, BB is rather uniformly low, and MSCI concentrates on mid-high values (Appendix C, Panel A). For ESG ratings, the distribution for REF is similar, while BB and MSCI scores are more centered, although BB scores are lower on average (Appendix C, Panel B).

Panel A: ENV raw score difference



Panel B: ENV raw score difference (sample without zero rating observations)

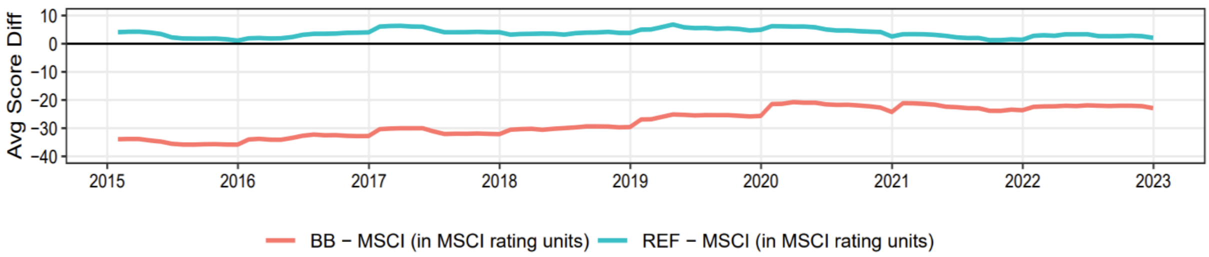
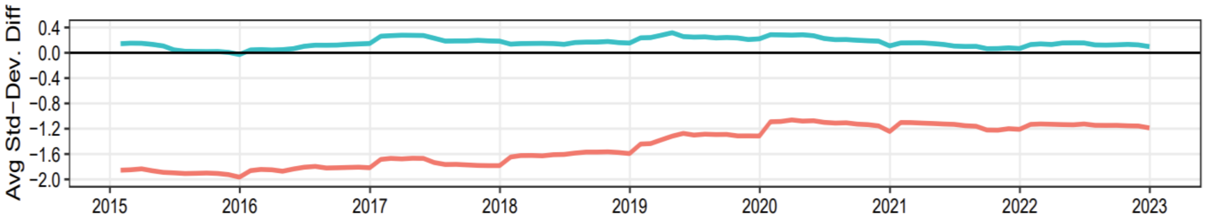


Figure 2. Average difference in index's environmental ratings across providers, January 2015–December 2022

[20] The statistics of ENV and ESG ratings show that REF is the most generous of all three providers. BB ENV and ESG average ratings are markedly lower than those assigned by both MSCI and REF (Appendix B).
[21] The share of ENV scores with a value of zero is 3 %, 14.2 %, and 0.3 % for REF, BB, and MSCI. Due to the comparatively low weights those companies receive in the indices, the exclusion of zero-rating firms does not drastically change the mean index scores (Appendix B, Panel A).
[22] We find qualitatively similar results for the differences in ESG ratings issued by REF, BB, and MSCI.

Panel C: ENV score difference measured in unit of MSCI ENV score Std-Dev.



Panel D: ENV score difference measured in unit of MSCI ENV score Std-Dev. (sample without zero rating observations)

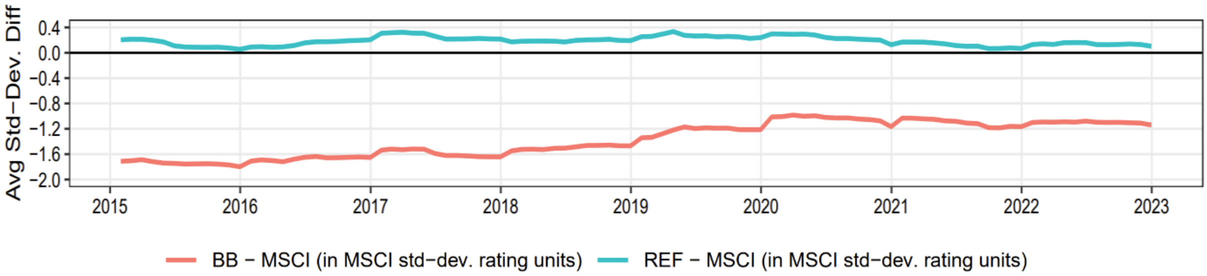


Figure 2. Average difference in index's environmental ratings across providers, January 2015–December 2022 (continued)

This figure presents the average monthly difference in an index's firm-weighted environmental (ENV) ratings between Bloomberg (BB) and MSCI and between Refinitiv (REF) and MSCI. For each month, we calculate the ENV rating of each index in each region, measured as firms' ratings weighted by firms weights. We then average the index's monthly ratings across all four regions (the World, U.S., Europe, and Emerging Markets) and across all five examined indices namely the MSCI Standard Index, the MSCI Climate Change Index (CC), the MSCI Climate Paris-Aligned Index (PAB), the MSCI Socially Responsible Investment Index (SRI) and the MSCI SRI Filtered PAB Index (SRI PAB). The Refinitiv score ranges between 0 and 100, while the Bloomberg score ranges between 0 and 10. For ease of comparability, the Bloomberg values are scaled by a factor of 10. A higher score indicates better environmental performance assessment. Variable definitions are provided in Appendix I.

Panel A and B respectively show the average differences in the raw rating scores for the full sample and for the sample that excludes firms with zero ENV rating values. Panel C and D respectively feature the difference in the rating scores (between BB and MSCI and between REF and MSCI) expressed as the numbers of the standard deviations (std dev) of MSCI ratings.

Untabulated analysis of the full sample shows a noticeable divergence between BB and MSCI (BB-MSCI) ratings, with the average difference of -30.1 points or -1.48 std dev of MSCI ratings, and ranging between -39.7 and -22.5 (-1.98 and -1.06 std dev of MSCI ratings). In contrast, REF and MSCI ratings are quite similar, with the average difference (REF-MSCI) of merely 3.12 points or 0.157 std dev of MSCI ratings, and varying between -0.78 and 6.42 (-0.032 and 0.317 std dev of MSCI ratings).

3.2 Method

3.2.1 Index's environmental performance measures

Similar to an approach employed by Abouarab et al. (2022), Humphrey and Li (2021), and Rohleder et al. (2022) to calculate a fund's carbon footprint, we compute the value of an environmental performance measure (EPM) of index *i* at time *t* as the weighted sum of firm *j*'s weight ω in index *i* at time *t* and firm *j*'s *EPM* value, as follows:

$$EPM_{i,t} = \sum_{j=1}^{J} \omega_{i,j,t} * EPM_{j,t} \tag{1}$$

where:

EPM is either total carbon emissions (Scope 1+2, the primary measure), or alternatively, Scope 1 emissions, carbon intensity, ESG/Environment rating scores, as outlined in the previous section.

J is the total number of firms in an index *i* at time *t*.

The EPM is calculated on a monthly basis to make use of the monthly constituent weighting data. A common challenge is that CO_2 emissions and ESG/environment rating are available at annual frequency. To compute an index's firm-weighted rating score for a month, each constituent's rating score is repeated at monthly frequency. To calculate an index's firm-weighted carbon emissions for a month, emissions of each constituent are first divided by 12.[23] The firm-weighted estimation provides us with an aggregate value of index *i*'s environmental performance in which each company *j*'s contribution is proportional to its weight ω in index *i*.

To address the influence of firm size on firm *j's* carbon emissions, we then introduce an adjusted variant of equation (1). This involves scaling the environmental performance *EPM* (carbon emissions) of index *i* by its firm-weighted market value at time *t* to normalize for differences in sizes. The adjusted environmental metric (AEPM) for index *i*'s CO_2 emissions at time *t* is estimated in tonnes per U.S. dollar of market capitalization, as follows:

$$AEPM_{i,t} = EPM_{i,t} * \frac{1}{\sum_{j=1}^{J} \omega_{i,j,t} * MV_{j,t}} \tag{2}$$

where $MV_{j,t}$ is the Market Value of a firm j at time t.

By estimating both unadjusted and size-adjusted carbon emissions, we can more accurately understand the "true" environmental performance (emissions footprint) of an index. Market value (instead of sales) is chosen to scale carbon footprint, as it allows us to directly map carbon emissions to valuation, which is of direct interest to investors.[24] For an alternative emissions measure used in a robustness test - *Carbon intensity*, we do not adjust it by size, as by definition, it is already normalized by either net sales (REF, BB) or total revenue (CDP). For ESG and

[23] CO_2 emissions may fluctuate throughout a year and the emissions pattern varies across 11 sectors in our sample and varies across time, particularly during the COVID lockdown period. Our simplified computation of monthly CO_2 footprint allows us to analyse an index's cumulative emissions over time (months). Other studies, such as Abouarab et al. (2022), employ lower frequency data, using quarterly intervals for the carbon emissions variable. A related study, Ulussever et al. (2023), uses monthly and yearly data to estimate carbon emissions, and suggests that the estimation performance of its models increases with high-frequency (i.e., monthly) data.

[24] An added benefit of using market value instead of sales is that it is available at a monthly frequency.

ENV rating scores, they are not assigned based on company size[25]; thus, the raw rating scores are used in the analysis.

The reduction in carbon footprint associated with a green index over time is of great importance to investors who care about sustainability. Thus, we also calculate the cumulative AEPM for total (scope 1 + 2), and its two components, direct (scope 1) and indirect (scope 2) emissions, given a US$1000 investment at the beginning of the study period. The cumulative sum of size-adjusted CO_2 emissions in kilograms (kg) is obtained for each index i over the period January 2015–December 2022.

3.2.2 OLS index-level regressions

Index-level OLS analyses were conducted to investigate the effect of an index's Green Criteria Effectiveness (GCE) on its environmental performance. For each index-month observation, we control for index's constituents' size-, sector-, and country characteristics. As large firms, all else equal, are associated with greater carbon emissions, we account for the size effect by computing an index's firm-weighted Market Value in each month. Since firms in developed countries tend to have higher emission standards, the Human Development Index (HDI), which reflects the development level of a constituent's country of residence, is used to account for heterogeneity across countries in the sample.[26] Sector composition of the index is also controlled for, since sectors such as Energy are prone to higher emissions.[27] As constituents' weights change each month and constituents in an index also change throughout the year, an index's values for size, sector, and HDI change each month. A Post-Launch dummy variable is introduced to account for the fact that some indices' data are backtested by MSCI for a sub-period during our study. Finally, controls for the COVID-19 pandemic and the Ukraine war are included, as these events affected firms' emissions through lockdowns and supply chain disruptions. These control variables allow us to isolate the effect of the green criteria MSCI

[25] Although larger companies tend to receive better ESG scores, market capitalization is not used as an influencing criterion in the construction of each rating score.

[26] Country-fixed effects are not used as there are 50 countries in the sample. Using country dummy variables will lead to high variance inflation factors (VIFs) in the OLS regression. To compute an index's (lagged) HDI value, countries in the sample are categorized into five groups based on their numeric HDI values in each year (≤ 0.6, 0.6–0.7, 0.7–0.8, 0.8–0.9, ≥ 0.9). The reference group is ≥ 0.9. The HDI value of firm j at time t is equal to the HDI value of the country where firm j is domiciled at time t. If the HDI value is not available for a year, the most recent year's value is taken. The index's value for an HDI group is the sum of firms' weights in that HDI group. If 15 % of the firms in an index have a HDI value between 0.7 and 0.8, the value of the HDI group (0.7–0.8) is 15 %. HDI data are gathered from the United Nations Development Programme (UNDP).

[27] The eleven sectors include Health Care, Consumer Discretionary, Materials, Industrials, Information Technology, Financials, Consumer Staples, Energy, Utilities, Communication Services, and Real Estate which is used as the reference category.

used to select constituents which may result in carbon emissions reduction compared to the standard index.

For each environmental performance metric discussed earlier, the following index-level OLS regression is estimated:

$$Y_{i,t} = \alpha + \sum_{k=1}^{4} \beta_i \times GCE_{Index_{ik}} + \sum_{g=1}^{4} \gamma_g \times HDI_{Group_{g,t-12}} + \sum_{s=1}^{10} \delta_s \times Sector_{s,t-1} \tag{3}$$

$$+\phi_1 Post\_Launch_{i,t-1} + \phi_2 lMV_{i,t-1} + \phi_3 Covid_{t-1} + \phi_4 Ukraine_{t-1} + \epsilon_t$$

where:

$Y_{i,t}$: An environmental performance measure for index i at time t.

GCE_Index_i : Dummy variable equal to 1 if index i belongs to a specific green index type, and 0 otherwise. For example, GCE_CC =1 for the CC index, and GCE_CC = 0 for all other indices (Standard, PAB, SRI, SRI PAB), and so on.

$HDI\_Group_{g,t-12}$: Percentage of firms in an HDI group g at time $(t-12)$.

$Sector_{s,t-1}$: Percentage of firms in sector s at time $(t-1)$.

$Post\_Launch_{i,t-1}$: Dummy variable indicating if index i had already been launched prior to or was launched at time $(t-1)$.

lMV_{t-1}: Logarithm of firm-weighted market value of index i at time $(t-1)$.

$Covid_{t-1}$: One-month lagged dummy variable for the COVID-19 period.

$Ukraine_{t-1}$: One-month lagged dummy variable for the Ukraine war period.

ϵ_t: Error term at time *t*.

In contrast to Popescu et al. (2023) which uses a dummy variable such as SRI and SFDR Article 9 to examine the effect of generic fund type on a fund's carbon intensity/carbon footprint, this study does not pool various green indices in one category. Instead, we treat four green index types separately and create four dummy variables GCEs accordingly. In equation (3), the standard index's GCE serves as the reference category and is subsumed into the intercept (α). A statistically significant and positive coefficient estimate for GCE_Index_i ($\beta_i > 0$) indicates its positive effect on index i's performance measure $Y_{i,t}$, after accounting for control variables.

Our primary performance measure is cumulative size-adjusted total carbon emissions (Scope 1 + 2). Cumulative CO_2 emissions reflect the total emissions over time, which aligns with the long-term perspective of investors and is central to investors' decision-making. While an OLS regression is not optimal to analyse cumulative measures, it is used for two main reasons. First, it provides clear and economically interpretable coefficients. This is important in this context where each dummy variable of interest, GCE, reveals CO_2 emissions reduction associated with investing in a green index, after accounting for the index's constituentweighted relevant variables. Second, Angrist and Pischke (2009) suggest that complicated econometric methods are often not needed and can even lead to biased estimates. The analyses of cumulative and non-cumulative performance measures show no significant difference, suggesting that using cumulative values does not alter the insights or the key findings, and hence supporting our OLS model choice.

We account for serial correlation and heteroskedasticity by estimating all models with the Newey-West standard errors, which address these issues.[28] The baseline OLS analyses of cumulative measures are complemented by a number of robustness OLS analyses using alternative environmental performance measures. Analyses utilizing samples of sub-indices, a sub-region, and market uncertainty sub-period are also performed. We obtain very similar findings, which enhance the robustness of our baseline results.

A potential concern in our analysis is reverse causality, as MSCI uses emissions and ratings to construct its green indices. It should be emphasized that MSCI's environmental performance metric is not employed in any regression. All control variables are lagged by a month (except HDI composite, by a year) in the baseline OLS analysis. In robustness analyses, samples with control variables lagged by either two months or three months (except HDI composite, by a year) are employed. Similar results are obtained, which eliminate concerns of reverse causality. It is also worth emphasizing that the very small generalized variance inflation factor (GVIF) for the group of four GCE dummy variables, estimated using DescTools in R, indicates no concern of multicollinearity in the analysis.

[28] We also estimate all models using the White standard errors and obtain similar results.

4 Empirical results

4.1 Carbon emissions analysis

The environmental performance of MSCI's green indices, relative to their standard indices, is evaluated using emissions data from REF, BB, and CDP, and rating scores issued by REF and BB.[29] To minimize reverse causality concerns, MSCI-issued emissions and rating scores are not used in any analysis.

Table 1 reports the cumulative size-adjusted CO_2 emissions (Scope 1 + 2) of each index in each region at three investment horizons, given a US$1000 investment at the beginning of the study. There are significant differences among the four regions. Over the study period, the cumulative emissions are highest (smallest) in EM (the U.S.), and the emissions differential between a green index and the standard index is substantially larger in EM than other regions.[30]

An investor could achieve significant CO_2 savings over time by investing in a green index instead of the standard index. As time passes, the benefit of an investment in a green index, measured in terms of CO_2 emissions savings, increases sharply. This illustrates the importance of a long-term investment perspective in such green equity products, as a significant amount of CO_2 emissions is avoided each year.

For a US$1000 investment in the two Paris Agreement-aligned PAB and SRI PAB indices (instead of the standard index) in All Regions (Panel A), an investor would have saved an average of 1429 (− 76.54 %) and 1355 (− 72.58 %) kg (column 3), 1253 (− 75.39 %) and 1170 (− 70.40 %) kg (column 6), 1256 (− 69.20 %) and 1143 (− 62.98 %) kg (column 9) total CO_2 emissions during the entire period, depending on the emissions provider which is REF, BB, and CDP respectively. This suggests that companies included in these green indices, on average, have lower carbon footprints, aligning with the environmental aims of these indices. This savings pattern holds for each individual region (Panel B–E). The lowest savings can be realized in the U.S. whereas the highest is in EM.[31]

[29] The emissions data from REF cover between 82% and 86% of constituents in the sample; those from BB, depending on the type of emissions, are between 56% and 84%, and for CDP emissions, between 50% and 70%.

[30] We find similar results, with the standard index having the highest emissions, followed by CC, SRI, PAB, and SRI PAB when using non-cumulative unadjusted or logarithm of non-cumulative unadjusted CO_2 emissions.

[31] These estimates are calculated as (438-1.867)/1.867 and (512-1.867)/1.867 kg for REF (Column 3, Panel A); (409-1.662)/1.662 and (492-1.662)/ 1.662 kg for BB (Column 6, Panel A); and (559-1.815)/1.815 and (672-1.815)/ 1.815 kg for CDP (Column 9, Panel A).

Table 1. Cumulative Scope 1+2 CO_2 emissions (in kg) given US$1000 investment at the beginning of the study, January 2015

Panel A: All Regions

	Refinitiv (REF)			Bloomberg (BB)			Carbon Disclosure Project (CDP)		
All regions	3 Years (1)	5 Years (2)	Total (3)	3 Years (4)	5 Years (5)	Total (6)	3 Years (7)	5 Years (8)	Total (9)
Standard	789	1336	1867	711	1149	1662	715	1176	1815
CC	396	645	885	353	561	788	406	639	928
PAB	248	359	438	228	336	409	307	456	559
SRI	366	570	720	328	498	650	357	567	761
SRI PAB	243	374	512	204	331	492	259	426	672
Panel B: World									
World	3 Years	5 Years	Total	3 Years	5 Years	Total	3 Years	5 Years	Total
Standard	490	732	967	439	666	886	476	729	1005
CC	266	392	509	228	342	448	260	393	530
PAB	140	193	237	113	157	191	135	191	235
SRI	300	450	558	267	398	498	313	454	560
SRI PAB	169	260	340	120	196	270	162	247	331
Panel C: The U.S.									
U.S.	3 Years	5 Years	Total	3 Years	5 Years	Total	3 Years	5 Years	Total
Standard	363	539	704	321	489	642	304	466	655
CC	174	257	333	152	229	299	148	227	314
PAB	130	190	221	111	167	194	94	138	169
SRI	183	290	371	157	255	324	172	261	328
SRI PAB	139	209	261	114	176	223	104	158	202
Panel D: Europe									
Europe	3 Years	5 Years	Total	3 Years	5 Years	Total	3 Years	5 Years	Total
Standard	586	905	1273	653	982	1343	687	1050	1474
CC	347	532	735	371	554	747	402	606	832
PAB	187	274	351	191	275	340	234	335	416
SRI	409	615	749	406	577	707	455	671	818
SRI PAB	179	284	406	101	186	302	208	317	454
Panel E: Emerging Markets									
EM	3 Years	5 Years	Total	3 Years	5 Years	Total	3 Years	5 Years	Total
Standard	1717	3165	4524	1430	2459	3777	1394	2459	4126
CC	796	1398	1964	660	1117	1660	813	1330	2035
PAB	534	781	942	498	746	908	765	1159	1415
SRI	573	926	1203	483	762	1072	490	881	1337
SRI PAB	486	743	1041	481	765	1171	560	983	1701

This table presents the cumulative size-adjusted total CO_2 emissions (in kg) associated with a US$1000 investment in each index in each region at the beginning of the study, January 2015. The total CO_2 emissions are adjusted based on an index's firm-weighted market value and the US$1000 initial investment. The variable definition (CO_2 total emissions Scope 1+2) can be found in Appendix I. The emissions data for All Regions, the World, the U.S., Europe, and Emerging Markets are reported by three providers, namely, Refinitiv (REF), Bloomberg (BB), and Carbon Disclosure Project (CDP). For brevity, we only report the total CO_2 emissions (Scope 1 + 2) at three investment horizons starting from the study beginning in January 2015. The three-year horizon ended in January 2018. The five-year horizon ended shortly before the spread of COVID-19 in January 2020. The total horizon covers the period over which emissions data is available, January 2015-December 2022.

The five indices consist of the MSCI Standard (parent) Index, the MSCI Climate Change Index (CC), the MSCI Climate Paris-Aligned Index (PAB), the MSCI Socially Responsible Investment Index (SRI) and the MSCI SRI Filtered PAB Index (SRI PAB).

Figure 3 depicts the total CO_2 emissions (Scope 1 + 2), the direct CO_2 emissions (Scope 1), and the indirect CO_2 emissions (Scope 2) in kilograms reported by REF, BB, and CDP, assuming a US$1000 investment in each index at the beginning of the study. Figure 3 complements Table 1; both show that all green indices are associated with substantially lower carbon emissions over time and regions, and this holds for all three emissions providers. Of particular interest is that the steepness of the slope associated with a green index's emissions for Scope 1 + 2 and Scope 1 decreases as time passes.

Next, an OLS index-level analysis of cumulative size-adjusted CO_2 emissions (Scope 1 + 2, Scope 1) and CO_2 intensity, sourced from Refinitiv (the primary provider), BB, and CDP, was performed. The results are reported in Table 2.

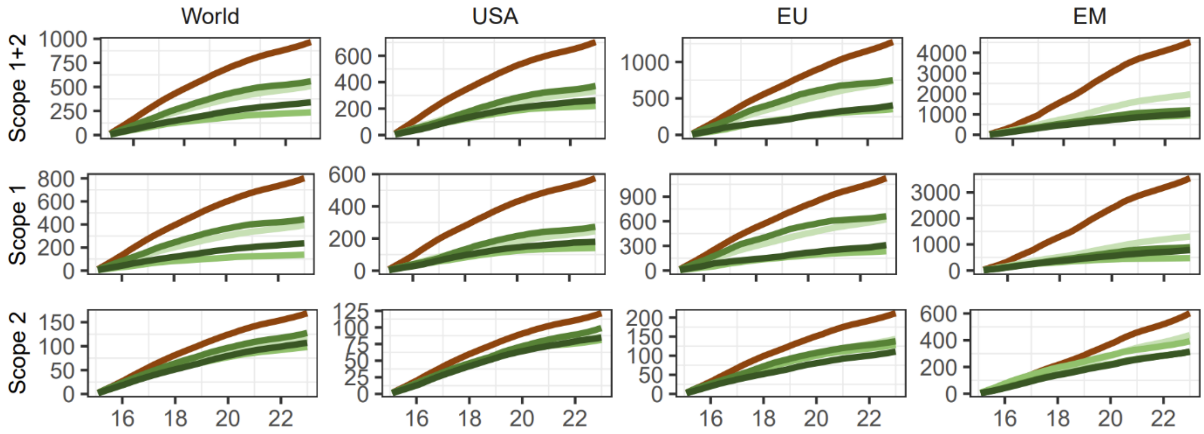
The analyses utilize cumulative size-adjusted total emissions (Scope 1 + 2, columns 1–3), cumulative size-adjusted direct emissions (scope 1, columns 4–6), their logged values (Scope 1 + 2, columns 7–9; Scope 1, columns 10–12), and emissions intensity (columns 13–15) from the three providers. Across the three providers and emissions measures employed, we find statistically significant and negative (at the 1 % level) GCE coefficient estimates for all green indices. The GCE PAB and GCE_SRI PAB have the strongest effect on total and direct emissions. For the REF emissions data, with a US$1000 investment at the beginning of the study, the GCE PAB (GCE_SRI PAB) lowers the total and direct emissions by 615.5 kg (587.6 kg) and 570.3 kg (532.6 kg) CO_2, respectively (columns 1 and 4).

The robustness analysis utilizing BB/CDP Scope 1 + 2, Scope 1 emissions, and emissions intensity generates consistently significant results. There are marked differences in the coefficient estimates derived from carbon intensity analysis. This is due to the fact that CDP CO_2 intensity is scaled by total revenue (column 15) whereas REF (BB) CO2intensity is scaled by net sales or revenue (sales) (columns 13–14).[32] However, the emissions savings are statistically significant at the 1 % level regardless of the scaling units.

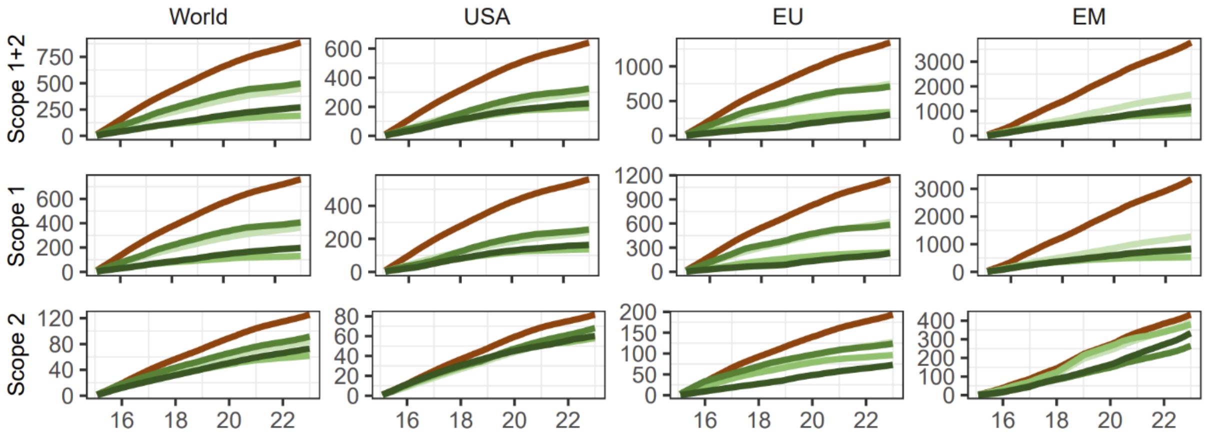
Overall, while Abouarab et al. (2022) and Popescu et al. (2023) report that not all of their analysed sustainable funds demonstrated emissions reductions, this study finds that the examined green indices, compared to the standard index, generate substantially lower CO_2 emissions. Our results are consistent with Jacob and Wilkens (2021); however, we emphasize the attractiveness of the PAB and SRI PAB indices that are aligned with the Paris Agreement, which are not examined by Jacob and Wilkens (2021).

[32] For definition of carbon intensity published by three providers REF, BB, CDP, please refer to Appendix I.

Panel A – Refinitiv**:** 1000$ Investment – Scope 1+2, 1, 2 in CO_2 Emissions (kg)



Panel B – Bloomberg: 1000$ Investment – Scope 1+2, 1, 2 in CO_2 Emissions (kg)



Panel C – CDP: 1000$ Investment – Scope 1+2, 1, 2 in CO_2 Emissions (kg)

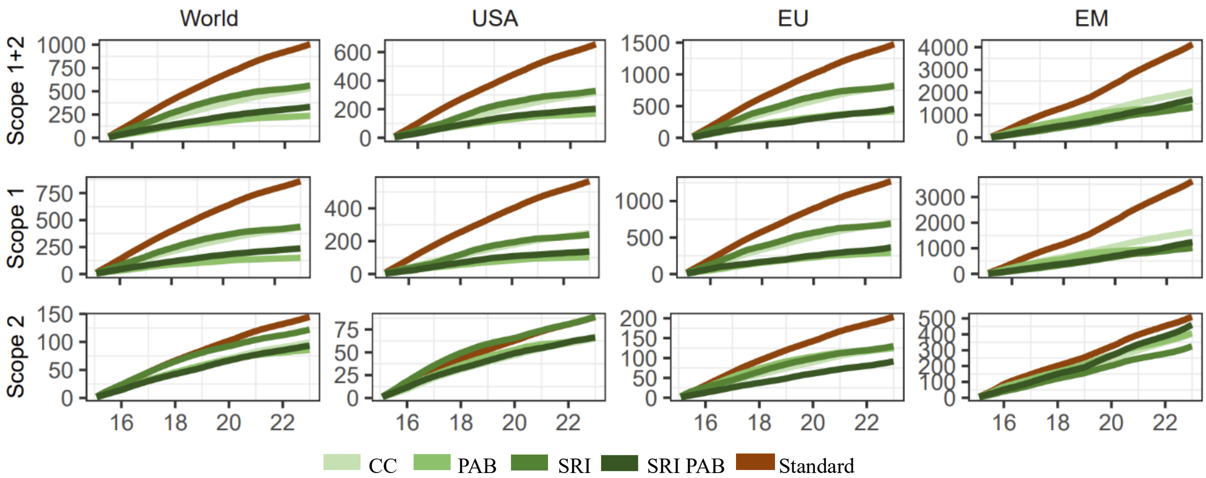


Figure 3. CO_2 emissions investment footprint, January 2015–December 2022

This figure shows the total CO_2 emissions, the direct CO_2 emissions, and the indirect CO_2 emissions (in kg) reported by REF, BB, and CDP, assuming a US$1,000 investment in each index at the beginning of the study. Each panel features the total Scope 1+2 CO_2 emissions, Scope 1 direct emissions, and Scope 2 indirect emissions across four regions. Variable definitions are provided in Appendix I.

The five examined indices consist of the MSCI Standard (parent) Index, the MSCI Climate Change Index (CC), the MSCI Climate Paris-Aligned Index (PAB), the MSCI Socially Responsible Investment Index (SRI) and the MSCI SRI Filtered PAB Index (SRI PAB).

Table 2. OLS regression analysis of cumulative size-adjusted CO_2 emissions (Scope 1+2, scope 1) and emissions intensity, January 2015–December 2022

Variables	Scope 1 + 2 Cumulative size-adjusted (Main measure)			Scope 1 Cumulative size-adjusted (Robustness test)			Scope 1 + 2 Log Cumulative size-adjusted (Robustness test)			Scope 1 Log Cumulative size-adjusted (Robustness test)			Scope 1 + 2 Emission Intensity (Robustness test)		
Provider	(1) REF	(2) BB	(3) CDP	(4) REF	(5) BB	(6) CDP	(7) REF	(8) BB	(9) CDP	(10) REF	(11) BB	(12) CDP	(13) REF	(14) BB	(15) CDP
GCE_CC	-303.9***	-274***	-209.1***	-300.2***	-285.3***	-206***	-0.4***	-0.5***	-0.4***	-0.5***	-0.5***	-0.4***	-51.7***	-38.1***	-2618.6***
	(36.1)	(28.0)	(30.2)	(26.5)	(23.7)	(25.9)	(0.05)	(0.06)	(0.05)	(0.06)	(0.06)	(0.06)	(4.5)	(3.2)	(125.8)
GCE_PAB	-615.5***	-532.7***	-441***	-570.3***	-532.7***	-428.9***	-1.4***	-1.4***	-1.3***	-1.7***	-1.7***	-1.5***	-86.2***	-51.3***	-3406.7***
	(56.6)	(43.1)	(48.7)	(39.4)	(35.7)	(41.4)	(0.11)	(0.11)	(0.11)	(0.11)	(0.11)	(0.1)	(6.0)	(4.3)	(157.1)
GCE_SRI	-366.2***	-349.2***	-310.4***	-319.4***	-339.8***	-304.9***	-0.4***	-0.4***	-0.3***	-0.4***	-0.5***	-0.4***	-72.6***	-42.7***	-3755.9***
	(33.3)	(28.9)	(28.5)	(29.3)	(26.9)	(26.2)	(0.04)	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)	(3.0)	(2.5)	(107.8)
GCE_SRI PAB	-587.6***	-575.7***	-514.5***	-532.6***	-546.1***	-491.4***	-0.7***	-1***	-0.7***	-0.8***	-1.1***	-0.8***	-63.7***	-68.6***	-4874.8***
	(50.6)	(44.7)	(44.4)	(42.7)	(38.9)	(39.6)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(5.3)	(3.9)	(163.1)
Sector composite	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
HDI composite	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Log of market value	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Post Launch dummy	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Ukraine dummy	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Covid dummy	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Observations	1900	1900	1900	1900	1900	1900	1900	1900	1900	1900	1900	1900	1900	1900	1900
Adj. R2	0.740	0.775	0.811	0.728	0.751	0.796	0.818	0.818	0.827	0.824	0.827	0.834	0.848	0.87	0.849

This table reports the coefficient estimates of the Green Criteria Effectiveness (GCE) of each green index obtained from the OLS index-level analysis of cumulative size-adjusted CO_2 emissions (Scope 1+2, Scope 1) and emissions intensity reported by Refinitiv (REF, Columns 1, 4, 7, 10, 13), Bloomberg (BB, Columns 2, 5, 8, 11, 14), and Carbon Disclosure Project (CDP, Columns 3, 6, 9, 12, 15). The dependent variable includes the cumulative size-adjusted total emissions (Scope 1+2, Columns 1, 2, 3), the cumulative size-adjusted direct emissions (Scope 1, Columns 4, 5, 6), the logarithm of cumulative size-adjusted total emissions (Scope 1+2, Columns 7, 8, 9), the logarithm of cumulative size-adjusted direct emissions (Scope 1, Columns 10, 11, 12), and CO_2 emission intensity (Columns 13, 14 and 15).

Refinitiv's emissions data is unscaled in tonnes, Bloomberg's original data is in thousands of tonnes, and CDP's original data in millions of tonnes. For ease of comparability, both Bloomberg and CDP' emissions data are scaled to tonnes. For columns 13-15, the different scaling of the emissions intensity results from the use of sales for REF and BB (columns 13 and 14) and the use of revenue for CDP (column 15). Variable definitions are provided in Appendix I.

Each model, estimated with the Newey-West robust standard errors, includes GCE index-type dummy values, lagged firm-weighted percentages for each Human Development Index (HDI) group, lagged firm-weighted percentages for each sector, logarithm of lagged firms-weighted market value, lagged dummy variables that capture an index's post-launch period, the Ukraine war and the COVID-19 pandemic. Index values are estimated using constituents' values and their weights in the index in each month, resulting in a total of 1900 observations derived from 20 indices and 95 monthly observations for each index. The five examined index types consist of the MSCI Standard (parent) Index, the MSCI Climate Change Index (CC), the MSCI Climate Paris-Aligned Index (PAB), the MSCI Socially Responsible Investment Index (SRI) and the MSCI SRI Filtered PAB Index (SRI PAB).
The coefficient estimates of GCEs are reported first, followed by the standard errors (in parentheses). For brevity, the coefficient estimates of the intercept (the standard parent index) and control variables are not reported. ***, ** and * represent p-value significance at the 0.01, 0.05, and 0.10 levels, respectively. The untabulated F-statistics for the tests of the joint significance of the coefficients of variables employed are significant at the 1% level in all model specifications

4.2 Robustness analysis: GCEs and carbon emissions

Robustness analyses (Appendix D) are performed utilizing alternative emissions measures including non-cumulative size-adjusted total emissions (Scope 1 + 2, columns 1–3), non-cumulative size-adjusted direct emissions (Scope 1, columns 4–6), logarithm of unadjusted total emissions (Scope 1 + 2, columns 7–9), and logarithm of unadjusted direct emissions (Scope 1, columns 10–12). The results presented in Appendix D are consistent with the results reported in Table 2, although the rankings in terms of the significant GCE magnitudes vary across models.

To eliminate concerns of reverse causality, we utilize a sample with two-month-lagged control variables and a sample with three-month-lagged control variables. For each sample, we perform OLS analyses of the four measures of REF total carbon emissions (Scope 1 + 2) including cumulative size-adjusted emissions, the logarithm of cumulative size-adjusted emissions, non-cumulative size-adjusted emissions, and the logarithm of non-cumulative unadjusted emissions. The results are reported in Appendix E. All GCEs impose consistently and statistically significant impacts (at the 1 % level), and their coefficient signs are as expected.

Next, OLS analyses utilizing various sub-samples are carried out. For brevity reasons, the results of four additional analyses are reported in Appendix F. For each sub-sample, the key environmental performance metric Scope 1 + 2 total emissions provided by REF, BB and CDP is analyzed. First, considering similar characteristics that CC index shares with PAB index, and SRI index shares with SRI PAB index, and to relax the assumption that the effect of each control variable is the same across indices, two OLS index-level regressions are performed: one for the standard, CC and PAB indices (columns 1–3) and one for the standard, SRI and SRI PAB indices (columns 4–6). Second, to mitigate a concern of serial correlation as there is overlap in country representation between the World region (dominated by the U.S.) and the U.S. region, an OLS index-level analysis of a sub-sample that does not include the five World indices (columns 7–9) is conducted. Third, a sub-sample consisting only of market uncertainty periods[33] (columns 10–12) is used. The results of these analyses remain consistent, and the coefficient signs of all significant GCEs are as expected.

Overall, across analyses, we obtain consistently and statistically significant (at the 5 % level) negative impacts of all green indices' GCEs on their total (Scope 1 + 2) carbon emissions.

[33] The periods of market uncertainty include the COVID-19 pandemic and the Ukraine war. Therefore, the cumulative variables, REF, BB, and CDP Scope 1 + 2 emissions, are recalculated starting in January 2020.

4.3 Additional analysis: GCEs and environmental/ESG rating scores

For additional tests, OLS index-level analyses of various rating scores[34] issued by REF and BB are performed, and the results are reported in Table 3. REF's ESG Controversies Score (column 1), REF's ENV score (columns 2–3), and REF's ESG score (column 6) are employed. As ESG ratings differ across providers, we additionally use BB's Environmental Score (columns 4–5) and BB's ESG scores (column 7).[35] REF and BB rating scores are independent of MSCI, which minimizes concerns of reverse causality. It should be noted that the REF and BB samples include a non-negligible number of observations with zero ENV rating scores. Thus, we report the results of the OLS analysis for ENV ratings using full samples (REF-column 2, BB-column 4) and samples without zero-rating observations (REF-column 3, BB-column 5).[36]

For the ENV Score and ESG Score issued by REF and BB, we document the statistically significant (at the 1 % level) and positive effects of the GCEs for the CC index and two most restrictive indices, SRI and SRI PAB. The baseline REF estimates show that GCE_SRI (GCE_SRI PAB) respectively improve REF's Controversies, ENV, and ESG scores by 9.8 (9.5), 4.44 (3.78) and 5 (4.74) rating points (see columns 1, 2 and 6).[37] However, the GCE_PAB estimates are significant and negative, albeit small, in the analysis of REF ENV (columns 2–3) and REF ESG scores (column 6). The subsequent analyses (Appendices G and H) will investigate this puzzling result.

For the BB ENV rating scores (columns 4–5), all four GCE coefficients are statistically significant and positive. The effect is strongest for GCE_SRI PAB followed by GCE_SRI. The former (the latter) respectively raise BB ENV rating by 9.8 and 6.2 points (column 4). The coefficients for GCE_CC and GCE_PAB are also significant and positive, but they have negligible effects, with increases of 3.7 and 1.2 points, respectively (column 4).

[34] These rating scores are only used in additional tests, as those ratings frequently emphasize policies and symbolic gestures over measurable reductions in environmental or social impacts, and are often unclear, unverified, and difficult to replicate with public data (Gonenc and Scholtens, 2017; Halbritter and Dorfleitner, 2015; Delmas et al., 2013).

[35] The rating data from REF covers 96 % of constituents in the sample. BB ratings covers 84 % for ENV and 95 % for ESG. MSCI ratings (not included in OLS regressions) covers 89 % for ENV and ESG in the sample.

[36] We do not observe a noticeable number of constituents with zero ESG rating scores

[37] The Refinitiv score ranges between 0 and 100, while the Bloomberg score ranges between 0 and 10. Bloomberg values were scaled by a factor of 10 to make them comparable with the Refinitiv values before we estimate regressions. For all scores (including the REF controversies score), a higher value indicates a better environmental performance assessment.

Table 3. OLS regression analysis of Refinitiv (REF) and Bloomberg (BB) rating scores, January 2015-December 2022

Variables	ESG Controversies score	Environmental (ENV) rating score				ESG rating score	
Provider	(1) REF	(2) REF	(3) REF0	(4) BB	(5) BB0	(6) REF	(7) BB
GCE_CC	-0.62	1.65***	1.81***	3.67***	3.9***	1.55***	1.07***
	(0.52)	(0.27)	(0.22)	(0.47)	(0.46)	(0.3)	(0.31)
GCE_PAB	1.84**	-2.22***	-1.67***	1.21*	2.2***	-1.89***	-0.45
	(0.78)	(0.48)	(0.39)	(0.71)	(0.66)	(0.54)	(0.53)
GCE_SRI	9.81***	4.44***	3.69***	6.23***	5.2***	5.03***	2.49***
	(0.31)	(0.2)	(0.18)	(0.36)	(0.32)	(0.21)	(0.25)
GCE_SRI PAB	9.47***	3.78***	3.15***	9.79***	9***	4.74***	4.21***
	(0.65)	(0.31)	(0.26)	(0.69)	(0.66)	(0.39)	(0.44)
Sector composite	YES	YES	YES	YES	YES	YES	YES
HDI composite	YES	YES	YES	YES	YES	YES	YES
Log of market value	YES	YES	YES	YES	YES	YES	YES
Post Launch dummy	YES	YES	YES	YES	YES	YES	YES
Ukraine dummy	YES	YES	YES	YES	YES	YES	YES
Covid dummy	YES	YES	YES	YES	YES	YES	YES
Observations	1900	1900	1900	1900	1900	1900	1900
Adj. R2	0.889	0.915	0.922	0.937	0.932	0.891	0.923

This table reports the coefficient estimates of the Green Criteria Effectiveness (GCE) of each green index obtained from the OLS index-level analysis of Refinitiv's (REF) ESG controversies rating score (column 1), REF's environmental (ENV) score (columns 2, 3), Bloomberg (BB)'s ENV score (columns 4, 5), REF's ESG score (column 6), and BB's ESG score (column 7). For ENV scores assigned by REF and BB, we report the analysis using the full sample (columns 2, 4) and using the sample that excludes zero ENV rating values (columns 3, 5). The Refinitiv score ranges between 0 and 100, while the Bloomberg score ranges between 0 and 10. For ease of comparability, the Bloomberg rating scores are scaled by a factor of 10 before we conduct each regression. A higher score (including controversies score) indicates better environmental performance assessment. Variable definitions are provided in Appendix I.

Each model, estimated with the Newey-West robust standard errors, includes GCE index-type dummy values, lagged firm-weighted percentages for each Human Development Index (HDI) group, lagged firm-weighted percentages for each sector, logarithm of lagged firms-weighted market value, lagged dummy variables that capture an index's post-launch period, the Ukraine war and the COVID-19 pandemic. Index values are estimated using constituents' values and their weights in the index in each month, resulting in a total of 1900 observations derived from 20 indices and 95 monthly observations for each index. The five examined index types consist of the MSCI Standard (parent) Index, the MSCI Climate Change Index (CC), the MSCI Climate Paris-Aligned Index (PAB), the MSCI Socially Responsible Investment Index (SRI) and the MSCI SRI Filtered PAB Index (SRI PAB).

The coefficient estimates of GCEs are reported first, followed by the standard errors (in parentheses). For brevity, the coefficient estimates of the intercept (the standard parent index) and control variables are not reported. ***, ** and * represent p-value significance at the 0.01, 0.05, and 0.10 levels, respectively. The untabulated F-statistics for the tests of the joint significance of the coefficients of variables employed are significant at the 1% level in all model specifications.

To understand the puzzling significant negative coefficient estimates of GCE_PAB in the analysis of REF ENV/ESG ratings (columns 2, 3, 6, Table 3), a Cumulative Weighting Difference (CWD) analysis (Appendix G) is carried out. This analysis examines the weighting differences between green indices and the standard index for constituents ranked by ENV ratings. The CWD results reveal that, when using REF and BB ratings, the PAB index tends to overweight companies with lower ENV scores and underweight constituents with higher scores, leading to an overall lower ENV rating compared to the standard index.

Additionally, an attribution analysis of each green index's ENV rating performance is performed (Appendix H). The environmental excess rating (between a green index and the standard index) is decomposed into sector allocation and stock selection components. For REF and BB ratings, the PAB and CC indices display negative excess ratings, with both sector allocation and stock selection contributing negatively. Specifically, the PAB index not only overweights sectors with lower environmental ratings but also selects constituents within sectors that have lower ratings compared to those in the standard index. In contrast, for MSCI ratings, all four green indices exhibit positive excess ratings, driven by both positive sector allocation and stock selection components. This suggests that MSCI's green index construction aligns with its environmental objectives.

The results of the CWD and attribution analysis suggest that the divergences in rating contents across providers contribute to the negative GCE effect observed for the PAB index.

4.4 Discussion and implication

This study offers practical implications for investors, index providers, fund managers and regulatory bodies. First, investors seeking to decarbonize should invest in ETFs or funds tracking the MSCI Paris Aligned Benchmark or SRI PAB. These options directly align with the Paris Agreement's goals and maintain broad market exposure. They also help lower portfolio carbon intensity. For example, a US$1000 investment in the Paris Aligned Benchmark, instead of the standard index, in January 2015 reduced about 1429 kg of CO_2 by December 2022.

Second, index providers and fund managers must clearly define their ESG objectives and performance measures and tie each fund to at least one sustainable development goal. They should adopt clear, consistent methodologies like MSCI. Transparent disclosures of environmental targets and measurement criteria allow investors to accurately assess the environmental impacts of green equity funds.

Third, policymakers and regulators should embed Paris Aligned Benchmarks into formal rules. They can grant lower capital charges to green benchmark products and require pension funds and insurers to hold a minimum share of assets in low-carbon indices. These measures will drive asset managers to adopt green benchmarks and speed the transition to a low-carbon economy.

Finally, the wide discrepancies among Refinitiv, Bloomberg, and CDP ratings show the urgent need for uniform standards. Regulators should require full disclosure of rating methods and promote harmonized metrics to make environmental claims consistent and comparable. This will boost investor confidence and reinforce trust in green finance.

5 Conclusion

This study examines the environmental performance of 20 MSCI equity indices across four regions over a relatively long period, 2015–2022. In each region, we analyse the MSCI Standard (parent) Index and four MSCI green indices, two of which are aligned with the Paris Agreement. Utilizing monthly constituent data, we investigate whether and how the GCE employed by MSCI in constructing each green index affects its environmental performance, after accounting for its firm-weighted size-, sector- and country-specific characteristics.

The analyses reveal two key findings. First, green indices were associated with substantially lower total CO_2 emissions (Scope 1 + 2 and Scope 1) than those of their respective standard indices. Thus, our results differ from Abouarab et al. (2022) and Popescu et al. (2023) as they did not document a reduction in carbon emissions for all of their analysed sustainable funds. Most importantly, we document that MSCI's GCEs used to construct green indices were statistically and significantly associated with indices' lowered total and direct emissions. The additional analyses utilizing ENV/ESG rating scores issued by REF and BB confirm the strong, significant positive effects of the GCEs for the two most restrictive indices, SRI and SRI PAB, but reveal a puzzling negative effect of the GCE for the PAB index. This puzzling result can be attributed to divergences in rating contents across providers, which suggest that employing rating scores assigned by different providers may lead to markedly different investment decisions.

This research makes several important contributions to the literature on sustainable finance. First, it provides new insights into the environmental performance of green indices relative to their standard counterparts, using a rich constituent dataset of MSCI indices with detailed monthly weighting data spanning from 2015 to 2022. Second, it is one of the first studies to analyse the PAB and SRI PAB indices, emphasizing their relevance in the context of the Paris Agreement-Aligned Benchmark investments. Third, this study enriches the discussion on the role of green equity investments as viable tools for reducing carbon emissions.

Our study complements the findings of Heldmann et al. (2025), which examines the financial performance of the same set of green MSCI indices relative to MSCI standard indices. Combining the results of both studies over the same period, we find that green MSCI indices outperform their parent indices both financially and environmentally, highlighting their appeal as an investment option for ethically minded investors and their relevance in accelerating the transition to a low-carbon economy.

This study is limited by its focus on 20 MSCI indices over a specific period that covers the pandemic and the ongoing war in Ukraine. The availability of historical constituent weighting data across regions is limited, and the study's findings may not be fully generalizable to other indices and/or other time periods. Furthermore, the analysis is constrained by the lack of a standardized framework for measuring environmental performance, which could affect the comparability of results using different metrics published by various providers. Future research could expand on this study by examining a broader set of indices, ideally covering a longer time period. Additionally, exploring the impacts of ESG criteria employed by various index providers on indices' environmental performance in various economic contexts could provide deeper insights.

References

Abouarab, R., Mishra, T., Wolfe, S., 2022. Spotting portfolio greenwashing in sustainable funds. SSRN Electron. J. https://doi.org/10.2139/ssrn.4258128.

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

Bajic, A., Kiesel, R., Hellmich, M., 2023. Handle with care: challenges in company-level emissions data for assessing financial risks from climate change. J. Clim. Finance 5, 100017. https://doi.org/10.1016/j.jclimf.2023.100017.

Berg, F., Fabisik, K., Sautner, Z., 2021. Is History repeating itself?: The (Un)Predictable Past of ESG ratings. European Corporate Governance Institute - Finance Working Paper No. 708/2020. https://doi.org/10.2139/ssrn.3722087.

Berg, F., Kölbel, J.F., Rigobon, R., 2022. Aggregate confusion: the divergence of ESG ratings. Rev. Finance 26 (6), 1315–1344. https://doi.org/10.1093/rof/rfac033.

Beyhan, M., Kennedy, W., 2025. The Role of Indexes in the Energy Transition. S&P Global. https://www.spglobal.com/en/research-insights/special-reports/look-forward/the-role-of-indexes-in-the-energy-transition. accessed 20.05.25.

Bodie, Z., Kane, A., Marcus, A.J., 2014. Investments, 10th Global. McGraw-Hill Education, New York, NY.

Busch, T., Johnson, M., Pioch, T., 2022. Corporate carbon performance data: quo vadis? J. Ind. Ecol. 26 (1), 350–363. https://doi.org/10.1111/jiec.13008.

D'Ecclesia, R.L., Morelli, G., Stefanelli, K., 2022. Are the green ETFs really green? Rapporto Tecnico No. 4, Dipartimento di Scienze Statistiche, Sapienza Università di Roma. https://www.dss.uniroma1.it/en/system/files/pubblicazioni/Stefanelli%2C%20Morelli%2C%20D%27Ecclesia%202.pdf accessed 22.05.25.

Dai, R., Liang, H., Ng, L., 2021. Socially responsible corporate customers. J. Financ. Econ. 142 (2), 598–626. https://doi.org/10.1016/j.jfineco.2020.01.003.

De Villiers, C., Jia, J., Li, Z., 2022. Corporate social responsibility: a review of empirical research using Thomson Reuters Asset4 data. Account. Finance 62 (4), 4523–4568. https://doi.org/10.1111/acfi.13004.

Delmas, M.A., Burbano, V.C., 2011. The drivers of greenwashing. Calif. Manag. Rev. 54 (1), 64–87. https://doi.org/10.1525/cmr.2011.54.1.64.

Delmas, M.A., Etzion, D., Nairn-Birch, N., 2013. Triangulating environmental performance: what do corporate social responsibility ratings really capture? Acad. Manag. Perspect. 27 (3), 255–267. https://doi.org/10.5465/amp.2012.0123.

Dumitrescu, A., Gil-Bazo, J., Zhou, F., 2022. Defining greenwashing. SSRN Electron. J. https://doi.org/10.2139/ssrn.4098411.

Flammer, C., 2021. Corporate green bonds. J. Financ. Econ. 142 (2), 499–516. https://doi.org/10.1016/j.jfineco.2021.01.010.

Gonenc, H., Scholtens, B., 2017. Environmental and financial performance of fossil fuel firms: a closer inspection of their interaction. Ecol. Econ. 132, 307–328. https://doi.org/10.1016/j.ecolecon.2016.10.004.

Halbritter, G., Dorfleitner, G., 2015. The wages of social responsibility – where are they? A critical review of ESG investing. Rev. Financ. Econ. 26 (1), 25–35. https://doi.org/10.1016/j.rfe.2015.03.004.

Hamill, L., Wu, I., 2023. The Critical Role of Fixed Income Indices in the Climate Transition. FTSE Russell. https://www.lseg.com/en/insights/ftse-russell/critical-role-fixed-income-indices-climate-transition. accessed 20.05.25.

Heldmann, J., Brückner, T., Dang, H.D., 2025. Financial returns of going green: evidence from MSCI indices. J. Asset. Manag. https://doi.org/10.1057/s41260-025-00404-4.

Humphrey, J.E., Li, Y., 2021. Who goes green: reducing mutual fund emissions and its consequences. J. Bank. Finance 126, 106098. https://doi.org/10.1016/j.jbankfin.2021.106098.

Jacob, A., Wilkens, M., 2021. What drives sustainable indices? A framework for analyzing the sustainable index landscape. SSRN Electron. J. https://doi.org/10.2139/ssrn.3874437.

Kartal, M.T., Kılıç Depren, S., Pata, U.K., Taşkın, D., Şavlı, T., 2024. Modeling the link between environmental, social, and governance disclosures and scores: the case of publicly traded companies in the Borsa Istanbul sustainability index. Financ. Innov. 10. https://doi.org/10.1186/s40854-024-00619-1. Article 80.

Kaustia, M., Yu, W., 2021. Greenwashing in mutual funds. SSRN Electron. J. https://doi.org/10.2139/ssrn.3934004.

Lublóy, Á., Keresztúri, J.L., Berlinger, E., 2025. Quantifying firm-level greenwashing: a systematic literature review. J. Environ. Manag. 373, 123399. https://doi.org/10.1016/j.jenvman.2024.123399.

Meyers, S.M., Ferrero-Ferrero, I., Muñoz-Torres, M.J., 2024. ARE sustainable funds doing the talk and the walk? An ESG score analysis of fund portfolio holdings. Int. Rev. Econ. Finance 93, 1526–1541. https://doi.org/10.1016/j.iref.2024.04.023.

Morningstar Manager Research, 2024. Global sustainable fund flows: Q1 2024 in review. Morningstar direct. https://asiaapi.morningstar.com/ods_images/Global_Sustainable_Fund_Flows_2024Q1.pdf accessed 20.05.25.

MSCI Inc., 2022. MSCI launches new climate action indexes for investors seeking net zero transition progress [Press release]. MSCI. https://www.msci.com/documents/10199/98361e94-6130-7ae9-62c0-5ae1953ec60f. accessed 20.05.25.

Papadopoulos, G., 2022. Discrepancies in corporate GHG emissions data and their impact on firm performance assessment. JRC Working Papers in Economics and Finance, 2022/12. https://doi.org/10.2139/ssrn.4197950.

Popescu, I.-S., Gibon, T., Hitaj, C., Rubin, M., Benetto, E., 2023. Are SRI funds financing carbon emissions? An input-output life cycle assessment of investment funds. Ecol. Econ. 212, 107918. https://doi.org/10.1016/j.ecolecon.2023.107918.

Rohleder, M., Wilkens, M., Zink, J., 2022. The effects of mutual fund decarbonization on stock prices and carbon emissions. J. Bank. Finance 134, 106352. https://doi.org/10.1016/j.jbankfin.2021.106352.

Suchman, M.C., 1995. Managing legitimacy: strategic and institutional approaches. Acad. Manag. Rev. 20 (3), 571–610. https://doi.org/10.2307/258788.

Ulussever, T., Kılıç Depren, S., Kartal, M.T., Depren, Ö., 2023. Estimation performance comparison of machine learning approaches and time series econometric models: evidence from the effect of sector-based energy consumption on CO_2 emissions in the USA. Environ. Sci. Pollut. Res. 30 (18), 52576–52592. https://doi.org/10.1007/s11356-023-26050-0.

World Health Organization, 2020. Archived: WHO timeline - COVID-19. World Health Organization accessed 20.05.25. https://www.who.int/news/item/27-04-2020-who-timeline---covid-19.

Appendix A. MSCI's process to construct green equity indices

Our time series data of twenty MSCI indices covers four different regions, which represent most of the traditional investment universe, namely the World, the U.S., Europe, and EM. In each region, we examine the standard index representing the parent index and four green indices, each being constructed based on the parent index with mild to severe exclusion criteria, resulting in 20 indices across four regions. The MSCI standard index does not require any screening or exclusion criteria. The four green indices constructed based on the standard index are the CC, PAB, SRI, and SRI PAB Index. These green indices differ primarily in their thresholds for exclusion criteria and the respective index construction method, as discussed below. Of the four green indices, the MSCI CC Index excludes the fewest index components, whereas the MSCI SRI Filtered PAB Index is constructed using the strictest exclusion criteria. Both the PAB and SRI PAB are directly aligned with the Paris Agreement's objectives.

A.1 The MSCI Climate Change (CC) Index, and the MSCI Climate PAB Index

For the MSCI CC Index (launched in June 2019) and the MSCI Climate PAB Index (launched in October 2020), the standard parent index (launched in March 1986, except for EM, in January 2001) is used as the starting universe.[38] A negative screening is then performed to exclude companies that either violate the social principles of the U.N. Global Compact or generate revenue from the production or distribution of controversial weapons, tobacco, and fossil fuels. The remaining constituents are then screened using a Low Carbon Transition (LCT) score, which is calculated by combining a company's carbon footprint, its climate-related risks and its ability to manage them. Companies with an LCT score of three are classified as "Solutions". These companies are given a higher weight in the final index than companies with lower LCT scores. In each of the five LCT categories, namely Solutions, Neutral, Operational Transition, Product Transition and Asset Stranding, a weighting adjustment is also made compared to the weights in the standard index. This results in the final non-excluded constituents with their respective weights in the CC Index. Considering the region, for example, with the MSCI Europe being the standard index, the above process generates the MSCI Europe CC Index.

For each region, based on the CC Index described above, stricter rules, in addition to the known rules aligned with the Paris Agreement, are applied to construct the Climate PAB Index. Companies included in the Climate PAB index must reduce their Weighted Average Carbon Intensity (WACI) by 50 % compared to the parent index. Furthermore, the carbon footprint

[38] The detailed methods of all MSCI indices can be found on the following website under the name of the respective index: https://www.msci.com/index-methodology.

must decrease by 7 % annually. Companies that generate profits from fossil fuel-related activities are also excluded. This means that companies involved in exploring and processing activities in coal, oil, natural gas, and electricity generation with an excessive greenhouse gas (GHG) intensity are excluded. Conclusively, the final weight of a company in the MSCI Climate PAB Index is determined using the LCT score.

A.2 The MSCI Socially Responsible Investing (SRI) Index, and the MSCI SRI Filtered PAB (SRI PAB) Index

In contrast to the MSCI CC Index discussed above, the MSCI SRI Index (launched in June 2011, except for EM, in March 2014) and the MSCI SRI PAB index (launched in June 2020) typically employ more exclusions as a broader range of controversial activities are defined and companies that violate an index principle or exceed the respective percentage thresholds are excluded. The process considers the profits generated with the help of activities defined as controversial. Table A1 outlines the detailed thresholds companies must meet in order to be included in these two indices.

The MSCI SRI PAB index is the strictest, and at the same time, complies with sustainable ESG investment and climate objectives. Stricter thresholds are set for constituents that derive revenues from certain oil and gas or thermal coal activities. In addition to the exclusion of constituents that engage in controversial and unsustainable activities, individual stocks with particularly severe violations of sustainable investment objectives are flagged red and receive an ESG Controversies Score of 0. The remaining companies are then assessed using the MSCI ESG Ratings, with the assigned scores descending from AAA to CCC. Existing index constituents must have a minimum MSCI ESG rating of BB and an ESG Controversies Score of 1 to be finally selected, whereas new holdings require a minimum MSCI ESG rating of A and an ESG Controversies Score of 4.

In the final step of index optimization, a best-in-class approach is applied, which means that only the best 25 % of companies from each sector are included. Similarly to the Climate PAB, companies included in the SRI PAB index must reduce their Weighted Average Carbon Intensity (WACI) by 50 % compared to the parent index. Furthermore, the carbon footprint must decrease by 7 % annually. The maximum weight of an individual stock in the MSCI SRI PAB Index is limited to 5 %, which prevents significant overweighting of any individual constituent.

Table A.1: Activity revenue thresholds

Activity		Index Thresholds (Max. % of revenues)		
Category	Subcategory	Standard	SRI	SRI PAB
Controversial weapons	-	-	0%	0%
Conventional weapons	Production	-	5%	5%
	Components	-	15%	5%
Civilian firearms	Production	-	0%	0%
	Distribution	-	5%	5%
Nuclear	Weapons	-	0%	0%
	Power	-	5%	5%
	Supplier	-	15%	5%
Thermal coal	Mining	-	0%	0%
	Power Generation	-	5%	5%
Tobacco	Production	-	0%	0%
	Distribution	-	5%	5%
Alcohol	Production	-	5%	5%
	Distribution	-	15%	15%
Gambling	Ownership	-	5%	5%
	Services	-	15%	15%
Adult Entertainment	Production	-	5%	5%
	Distribution	-	15%	15%
GMO	Production	-	5%	5%
Oil and Gas	Conventional	-	-	0%
	Unconventional	-	0%	0%
	Power Generation	-	-	30%

This table shows the share of revenues that a company must not exceed in the respective activity in order not to be excluded from the MSCI SRI or SRI PAB index. The SRI and SRI PAB indices differ in that stricter exclusions apply to the SRI PAB in certain activities. The detailed methods of all MSCI indices can be found on the following website under the name of the respective index: https://www.msci.com/index-methodology.

Appendix B. Average environmental (ENV) and ESG rating scores, January 2015–December 2022

Panel A	ENV ratings					
	(1)	(2)	(3)	(4)	(5)	(6)
Index	REF	REF0	BB	BB0	MSCI	MSCI0
Standard	66.9	67.9	34.1	36.8	60.9	61.1
CC	66.6	67.7	33.5	36.4	63.1	63.3
PAB	65.3	66.4	32.6	35.7	62.6	62.7
SRI	71.1	71.4	37.7	39.5	68.7	68.7
SRI PAB	70.2	70.4	36.1	38.1	69.2	69.3

Panel B	ESG ratings		
	(1)	(2)	(3)
Index	REF	BB	MSCI
Standard	68.8	39.3	49.8
CC	68.4	38.4	50.4
PAB	67.3	38.2	51.3
SRI	73.5	40.8	58.8
SRI PAB	72.4	40.1	58.4

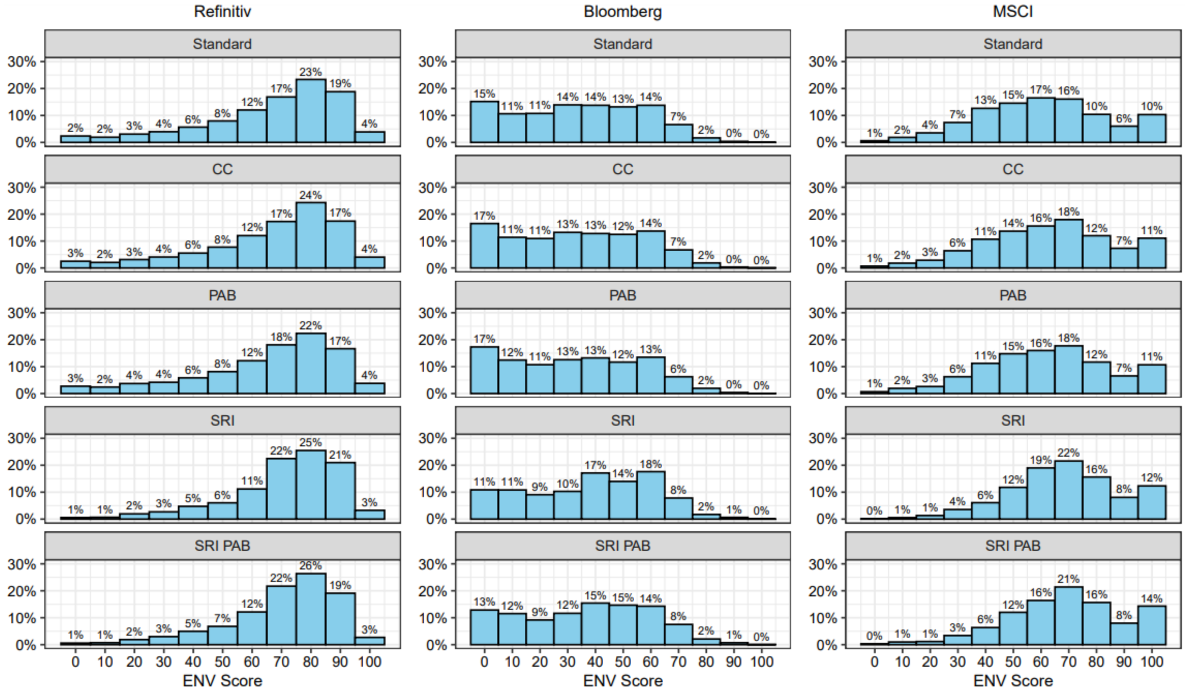
This Appendix presents the average ENV rating (Panel A) and average ESG rating (Panel B) of each index provided by Refinitiv (REF), Bloomberg (BB), and MSCI. For ENV scores, the statistics are reported using the full sample (columns 1, 3, 5) and using the sample that excludes zero ENV rating values (columns 2, 4, 6). The shares of ENV scores with a value of zero are 3%, 14.2% and 0.3% for REF, BB and MSCI. For ESG scores, we do not observe a sizable number of observations with zero ESG ratings; thus, we only report the statistics for the full sample. The Refinitiv scores range between 0 and 100, while the Bloomberg and MSCI scores range between 0 and 10. For ease of comparability, the Bloomberg and MSCI values are scaled by a factor of 10. A higher score indicates a better environmental performance assessment.

Rating data are available from the beginning of this study, January 2015, to December 2022. Variable definitions are provided in Appendix I. For each month, we calculate the ENV/ ESG rating of each index in each region, measured as firms' ratings weighted by firms' weights. The index's monthly ratings are averaged across all four regions (the World, U.S., Europe, and Emerging Markets).

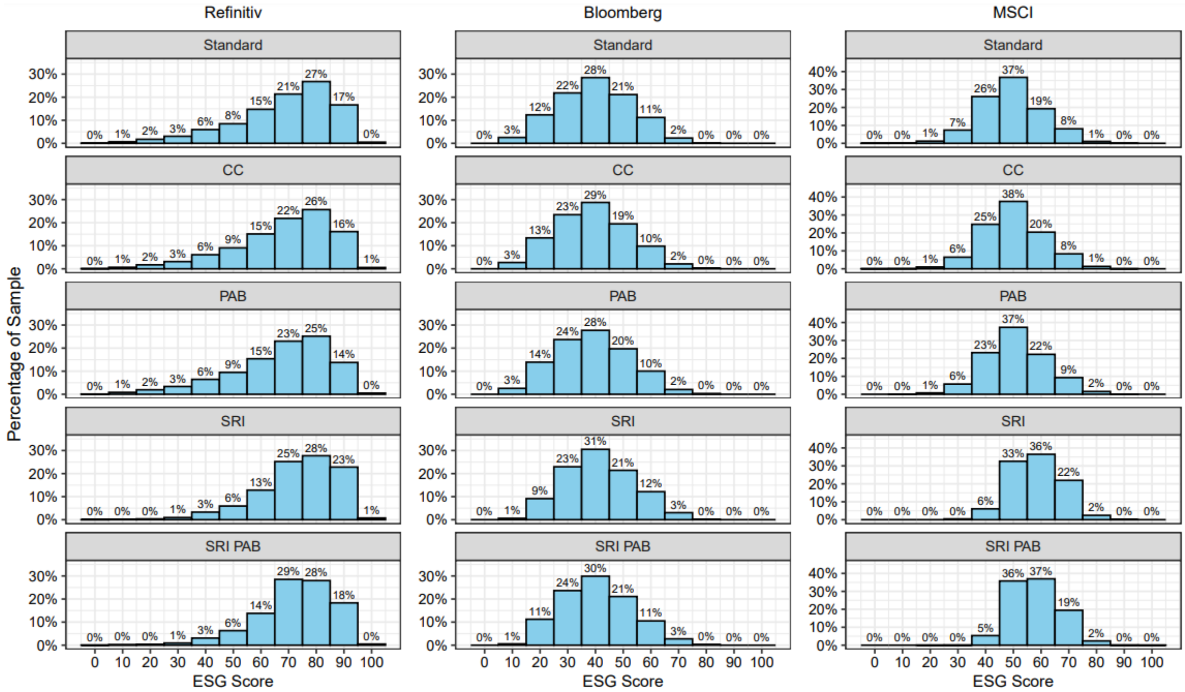
The five indices include MSCI Standard Index, the MSCI Climate Change Index (CC), the MSCI Climate Paris-Aligned Index (PAB), the MSCI Socially Responsible Investment Index (SRI) and the MSCI SRI Filtered PAB Index (SRI PAB).

Appendix C. Histogram of environmental (ENV) and ESG ratings, January 2015–December 2022

Panel A. Histogram of ENV rating scores



Panel B. Histogram of ESG rating scores



This appendix shows the histograms of environmental and ESG rating scores assigned by Refinitiv, Bloomberg, and MSCI for firms in the index weighted by their respective index weights across time and across regions. Panel A features the histograms for environmental scores. Panel B displays the histograms for ESG scores. The Refinitiv scores range between 0 and 100, while the Bloomberg and MSCI scores range between 0 and 10. For ease of comparability, the Bloomberg and MSCI scores are scaled by a factor of 10. A higher score indicates better environmental or ESG performance assessment. Variable definitions are provided in Appendix I. The five examined indices consist of the MSCI Standard (parent) Index, the MSCI Climate Change Index (CC), the MSCI Climate Paris-Aligned Index (PAB), the MSCI Socially Responsible Investment Index (SRI), and the MSCI SRI Filtered PAB Index (SRI PAB).

Appendix D. OLS robustness analysis of non-cumulative total and direct CO_2 emissions, January 2015–December 2022

Variables	Size-adjusted CO_2 emissions, Scope 1 + 2			Size-adjusted CO_2 emissions Scope 1			Logarithm of unadjusted CO_2 emissions, Scope 1 + 2			Logarithm of unadjusted CO_2 emissions Scope 1		
Provider	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	REF	BB	CDP	REF	BB	CDP	REF	BB	CDP	REF	BB	CDP
GCE_CC	-56.9***	-55.1***	-17.5***	-50.5***	-52.2***	-52.7***	-0.176***	-0.585***	-0.063	-0.197***	-0.145***	-0.294***
	(6.8)	(6.2)	(5.3)	(6.1)	(5.9)	(5.6)	(0.019)	(0.036)	(0.039)	(0.022)	(0.028)	(0.033)
GCE_PAB	-110.5***	-97.1***	-32.5***	-96.5***	-93.8***	-94.5***	-0.625***	-0.972***	-0.401***	-0.997***	-0.977***	-1.109***
	(8.9)	(7.5)	(7.6)	(7.6)	(7.3)	(6.9)	(0.025)	(0.049)	(0.048)	(0.03)	(0.036)	(0.042)
GCE_SRI	-102***	-101.4***	-53.1***	-87.6***	-94.8***	-91.6***	-0.681***	-0.95***	-0.41***	-0.884***	-0.924***	-0.914***
	(5.3)	(4.0)	(3.6)	(4.4)	(3.8)	(3.9)	(0.015)	(0.027)	(0.023)	(0.016)	(0.017)	(0.025)
GCE_SRI PAB	-81***	-94.4***	-62.5***	-72***	-84.4***	-88.5***	-0.974***	-1.539***	-1.026***	-0.994***	-1.071***	-1.1***
	(8.8)	(6.8)	(5.0)	(7.7)	(6.6)	(6.4)	(0.031)	(0.047)	(0.058)	(0.035)	(0.036)	(0.048)
Sector composite	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
HDI composite	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Log of market value	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Post Launch dummy	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Ukraine dummy	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Covid dummy	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Observations	1900	1900	1900	1900	1900	1900	1900	1900	1900	1900	1900	1900
Adj. R2	0.766	0.82	0.822	0.75	0.778	0.808	0.921	0.834	0.817	0.931	0.925	0.895

This appendix reports the coefficient estimates of the Green Criteria Effectiveness (GCE) of each green index obtained from the OLS index-level analysis of non-cumulative CO_2 emissions reported by Refinitiv (REF, Columns 1, 4, 7, 10), Bloomberg (BB, Columns 2, 5, 8, 11), and Carbon Disclosure Project (CDP, Columns 3, 6, 9 and 12) using the full sample. The dependent variable is the non-cumulative size-adjusted total CO_2 emissions (Scope 1+2, Columns 1-3), non-cumulative size-adjusted CO_2 direct emissions (Scope 1, Columns 4-6), the logarithm of non-cumulative un-adjusted total CO_2 emissions (Scope 1+2, columns 7-9), and the logarithm of non-cumulative unadjusted CO_2 direct emissions (Scope 1, columns 10-12). Refinitiv's emissions data are unscaled in tonnes, Bloomberg's original data are in thousands of tonnes, and CDP's original data are in millions of tonnes. For ease of comparability, both Bloomberg and CDP's emissions data are scaled to tonnes.

Each model, estimated with the Newey-West robust standard errors, includes GCE index-type dummy values, lagged firm-weighted percentages for each Human Development Index (HDI) group, lagged firm-weighted percentages for each sector, the logarithm of lagged firm-weighted market value, lagged dummy variables that capture an index's post-launch period, the Ukraine war and the COVID-19 pandemic. Index values are estimated using constituents' values and their weights in the index in each month, resulting in a total of 1900 observations derived from 20 indices and 95 monthly observations for each index. Variable definitions are provided in Appendix I.

The five indices include MSCI Standard Index, the MSCI Climate Change Index (CC), the MSCI Climate Paris-Aligned Index (PAB), the MSCI Socially Responsible Investment Index (SRI) and the MSCI SRI Filtered PAB Index (SRI PAB).

The coefficient estimates of GCEs are reported first, followed by the standard errors (in parentheses). For brevity, the coefficient estimates of the intercept and control variables are not reported. ***, ** and * represent p-value significance at the 0.01, 0.05, and 0.10 levels, respectively. The untabulated F-statistics for the tests of the joint significance of the coefficients of variables employed are significant at the 1% level in all model specifications.

Appendix E. OLS robustness analysis utilizing samples with two- and three month-lagged control variables, January 2015–December 2022

Sample	Sample with two-month lagged control variables				Sample with three-month lagged control variables			
	Refinitiv (REF) Scope 1 + 2 emissions				REF Scope 1 + 2 emissions			
Provider	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	CUM SA	LOG CUM SA	SA	LOG UA	CUM SA	LOG CUM SA	SA	LOG UA
GCE_CC	-275.8***	-0.386***	-56.7***	-0.199***	-258.2***	-0.347***	-56.2***	-0.216***
	(36.4)	(0.052)	(6.8)	(0.02)	(37.7)	(0.05)	(6.9)	(0.021)
GCE_PAB	-575.2***	-1.325***	-111.2***	-0.66***	-550.9***	-1.25***	-110.7***	-0.687***
	(54.8)	(0.099)	(8.9)	(0.027)	(54.5)	(0.092)	(9.0)	(0.028)
GCE_SRI	-359.7***	-0.336***	-104.1***	-0.695***	-363.3***	-0.333***	-105.1***	-0.703***
	(34)	(0.038)	(5.3)	(0.016)	(34.4)	(0.036)	(5.2)	(0.016)
GCE_SRI PAB	-569.5***	-0.681***	-79.7***	-0.985***	-567.7***	-0.661***	-76.4***	-0.99***
	(51.8)	(0.073)	(8.7)	(0.033)	(54.4)	(0.07)	(8.7)	(0.034)
Sector composite	YES	YES	YES	YES	YES	YES	YES	YES
HDI composite	YES	YES	YES	YES	YES	YES	YES	YES
Log of market value	YES	YES	YES	YES	YES	YES	YES	YES
Post Launch dummy	YES	YES	YES	YES	YES	YES	YES	YES
Ukraine dummy	YES	YES	YES	YES	YES	YES	YES	YES
Covid dummy	YES	YES	YES	YES	YES	YES	YES	YES
Observations	1880	1880	1880	1880	1860	1860	1860	1860
Adj. R2	0.74	0.834	0.768	0.919	0.743	0.846	0.772	0.918

This Appendix reports the coefficient estimates of the Green Criteria Effectiveness (GCE) of each green index obtained from the OLS index-level analysis of our four measures of total Scope 1+2 carbon emissions (in tonnes) published by REF (our main data provider), namely: cumulative (CUM) size-adjusted (SA) total emissions (CUM SA, columns 1, 5), logarithm (LOG) of cumulative (CUM) size-adjusted (SA) total emissions (LOG CUM SA, columns 2, 6), non-cumulative size adjusted (SA) total emissions (SA, columns 3, 7), and logarithm (LOG) of non-cumulative unadjusted (UA) total emissions (LOG UA, columns 4, 8).

We conduct this analysis utilizing a sample with two-month lagged control variables (1880 observations, columns 1-4) and a sample with three-month lagged control variables (1860 observations, columns 5-8), except index-level HDI composite, which is lagged by a year. Each model, estimated with the Newey-West robust standard errors, includes Green Criteria Effectiveness (GCE) dummy variables and lagged control variables. Index values are estimated using constituents' values and their weights in the index in each month. Variable definitions are provided in Appendix I.

The five indices include MSCI Standard Index (launched on 31 March 1986, except for EM, on 1 January 2001), the MSCI Climate Change Index (CC, launched on 20 June 2019), the MSCI Climate Paris-Aligned Index (PAB, launched on 26 October 2020), the MSCI Socially Responsible Investment Index (SRI, launched on 28 June 2011, except for EM, on 24 March 2014), and the MSCI SRI Filtered PAB Index (SRI PAB, launched on 26 June 2020).

The coefficient estimates of GCEs are reported first, followed by the standard errors (in parentheses). For brevity, the coefficient estimates of the intercept and control variables are not reported. ***, ** and * represent *p*-value significance at the 0.01, 0.05, and 0.10 levels, respectively. The untabulated F-statistics for the tests of the joint significance of the coefficients of variables employed are significant at the 1% level in all model specifications.

Appendix F. OLS robustness analysis utilizing various subsamples

Sample	Sample Standard, CC & PAB indices			Sample Standard, SRI & SRI PAB indices			Sample excluding "World" Region			Sample market uncertainty period		
Provider	(1) REF S1+2	(2) BB S1+2	(3) CDP S1+2	(4) REF S1+2	(5) BB S1+2	(6) CDP S1+2	(7) REF S1+2	(8) BB S1+2	(9) CDP S1+2	(10) REF S1+2	(11) BB S1+2	(12) CDP S1+2
GCE_CC	-163.9**	-175.4***	-264.6***	NA	NA	NA	-354.1***	-315.4***	-232.1***	-41.5***	-42.8***	-12.8
	(79.0)	(59.6)	(59.8)	NA	NA	NA	(45.2)	(35.2)	(37.4)	(18.9)	(16.5)	(13.3)
GCE_PAB	-550.1***	-477.8***	-541.8***	NA	NA	NA	-742.3***	-632.6***	-505.9***	-260.4***	-238.7***	-125.8***
	(113.4)	(84.8)	(87.4)	NA	NA	NA	(60.7)	(51.5)	(57.8)	(33.7)	(31)	(30.6)
GCE_SRI	NA	NA	NA	-449.2***	-404.9***	-356.6***	-392.3***	-376.3***	-330***	-156.8***	-138.2***	-51.5**
	NA	NA	NA	(51.5)	(41.5)	(43.3)	(44.1)	(34.7)	(35.3)	(26.2)	(23.8)	(24.1)
GCE_SRI PAB	NA	NA	NA	-955.5***	-859.9***	-774.9***	-610.7***	-605.5***	-524.4***	-159.3***	-159.3***	-124.3***
	NA	NA	NA	(91.7)	(74.1)	(78.3)	(61.0)	(51.8)	(52)	(29.9)	(26.6)	(21.5)
Sector composite	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
HDI composite	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Log of market value	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Post Launch dummy	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Ukraine dummy	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Covid dummy	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Observations	1140	1140	1140	1140	1140	1140	1425	1425	1425	700	700	700
Adj. R2	0.831	0.852	0.871	0.801	0.826	0.834	0.742	0.776	0.810	0.761	0.757	0.749

This Appendix reports the coefficient estimates of the Green Criteria Effectiveness (GCE) of each green index obtained from the OLS index-level analysis of the main environmental performance measure, cumulative size-adjusted total carbon emissions (scope 1+2 in tonnes) provided by Refinitiv (REF, columns 1, 4, 7, 10), Bloomberg (BB, columns 2, 5, 8, 11) and CDP (columns 3, 6, 9, 12).

This robustness analysis is conducted using four different sub-samples ending in December 2022. First, we examine the sub-sample that includes the standard index, and the two green indices that are less restrictive, the CC and PAB indices (columns 1-3). Second, we explore the sub-sample that consists of the standard index, and the two green indices that are most restrictive, the SRI and SRI PAB (columns 4, 5, 6). Third, we investigate a sub-sample that excludes five indices in the "World" region (columns 7, 8, 9). Finally, we examine the sub-sample that includes index-month observations during periods of market uncertainty (COVID-19 and the Ukraine war). For our last sub-sample (columns 10, 11, 12) our cumulative variables (Scope 1+2 emissions) are recalculated starting in January 2020.

Each model, estimated with the Newey-West robust standard errors, includes GCE dummy variables and control variables lagged by one month (except index-level HDI composite lagged by a year). Index values are estimated using constituents' values and their weights in the index in each month. Variable definitions are provided in Appendix I.

The five indices include MSCI Standard Index (launched on 31 March 1986, except for EM, on 1 January 2001), the MSCI Climate Change Index (CC, launched on 20 June 2019), the MSCI Climate Paris-Aligned Index (PAB, launched on 26 October 2020), the MSCI Socially Responsible Investment Index (SRI, launched on 28 June 2011, except for EM, on 24 March 2014), and the SRI Filtered PAB Index (SRI PAB, launched on 26 June 2020).

The coefficient estimates of GCEs are reported first, followed by the standard errors (in parentheses). For brevity, the coefficient estimates of the intercept and control variables are not reported. ***, ** and * represent *p*-value significance at the 0.01, 0.05, and 0.10 levels, respectively. The untabulated F-statistics for the tests of the joint significance of the coefficients of variables employed are significant at the 1% level in all model specifications.

Appendix G. Cumulative Weighting Difference (CWD) between green and standard indices for ENV rating-ranked constituents

G.1 CWD Method

An interesting question derived from the index's environmental rating analysis is whether constituents of higher rating scores are overweighted in an MSCI green index, relative to the MSCI standard index. To address this question, we first sort constituents in an index from the lowest to the highest ENV rating score assigned by a provider (REF, BB). Next, we calculate the difference at time t between a green index gi's weight $\omega_{gi,t}$ and the parent index pi's weight $\omega_{pi,t}$ for each rank. Then we obtain the cumulative sum of weighting differences (CWD) from the first ranked constituent up to the m-th ranked constituent:

An interesting question derived from the index's environmental rating analysis is whether constituents of higher rating scores are overweighted in an MSCI green index, relative to the MSCI standard index. To address this question, we first sort constituents in an index from the lowest to the highest ENV rating score assigned by a provider (REF, BB). Next, we calculate the difference at time t between a green index gi's weight $\omega_{gi,t}$ and the parent index pi's weight $\omega_{pi,t}$ for each rank. Then we obtain the cumulative sum of weighting differences (CWD) from the first ranked constituent up to the m-th ranked constituent:

$$CWD_{m,t} = \sum_{m=1}^{M} \left(\omega_{gi,m,t} - \omega_{pi,m,t}\right) \tag{4}$$

The CWD for each index in each region is plotted at random points in time (see Fig. G.1). In the plot it can be identified that the constituent whose rating is closest to the average rating score of the standard index, and is marked by a vertical red line there. A U shape with the bottom around the mean rating score (the tangent point between the red vertical line and the x axis) suggests that CWD decreases prior to reaching the average rating and increases after that.

G.2 CWD Results

The index-level rating analysis (Table 3) employs MSCI's constituent weights and constituents' ratings issued by REF and BB. The puzzling negative coefficient estimate of the GCE for PAB index (columns 2-3, 6, Table 3) suggests that MSCI PAB index's green criteria may not be well reflected in REF constituents' ratings which are used to derive PAB index's ratings. Motivated by this puzzling result, we examine the cumulative weighting difference (CWD) between a green index and the standard index for constituents ranked by ENV ratings issued by REF, BB, and MSCI. Fig. G.1 displays the CWDs across regions at three *random* points of time: January 2015 (Panel A), July 2019 (Panel B), and July 2022 (Panel C).

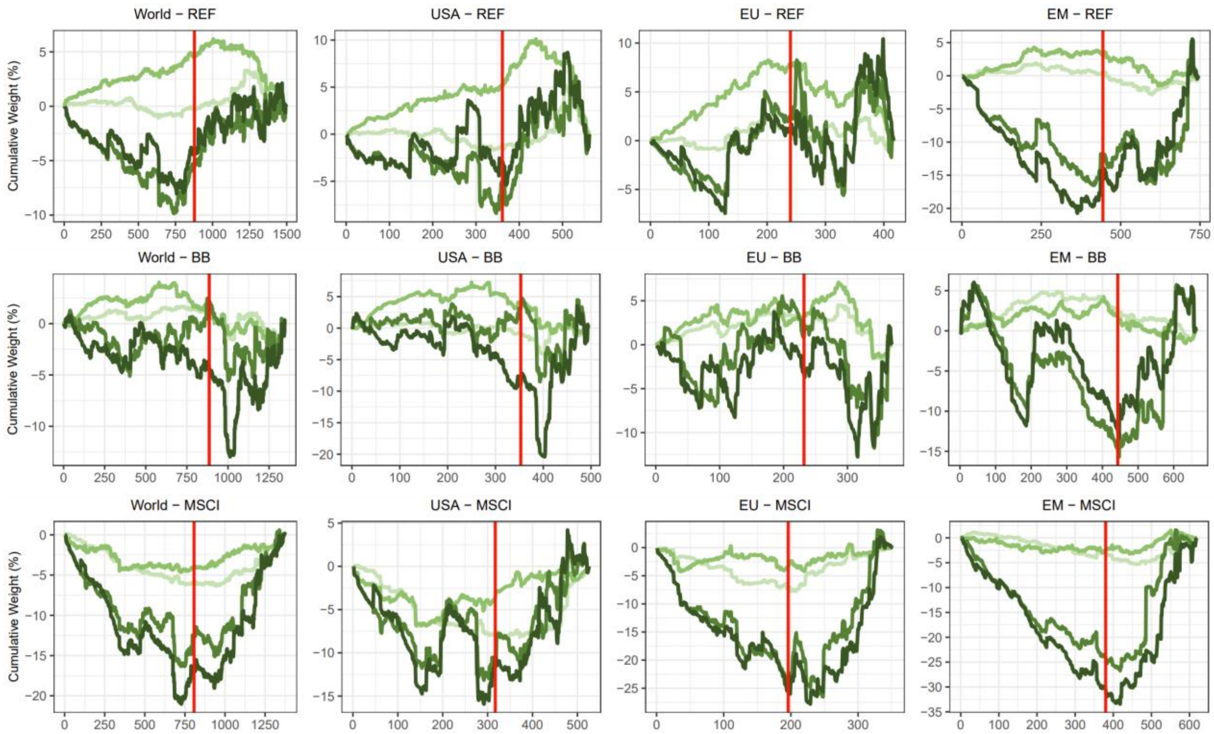
In each graph, the vertical red line marks the constituent whose ENV score is closest to the mean score of the standard index. If the CWD rises to the left of this red line, constituents with a relatively lower score are weighted more heavily in the green index than in the standard index, causing the overall score of the green index to fall. In contrast, if the CWD rises to the right of the red line, constituents with a relatively higher score are weighted more heavily in the green index than in the standard index, and the green index's overall score rises. If the CWD were flat, firms' weightings would be the same in both the green and the standard indices.

There are two striking features in MSCI's CWDs. First, two distinct pairs can be observed, CC and PAB, and SRI and SRI PAB; the latter exhibits a sharp decline in the CWD at the lower rating scores and a dramatic increase at the higher rating scores. Across three Panels corresponding to the three points of time, the CC and PAB have similar weightings, as do the SRI and SRI PAB. Second, except for the U.S., MSCI's CWD patterns are very clear, falling sharply at the scores preceding the average rating and then rising markedly, particularly in Panels B and C. This suggests that MSCI is "correctly" weighting the constituents according to their MSCI rating scores with an aim to achieve the desired higher (better) ENV score for each green index.

However, we do *not* observe consistently clear and sharp patterns in the CWDs constructed using MSCI's constituent weights *and* constituent ratings assigned by REF and BB, confirming a divergence in the contents of ratings assigned by the three providers. The patterns in the REF and BB CWDs are rather "messy". Across all regions, the PAB constituents are weighted higher at lower ENV scores and weighted lower at higher scores, suggesting a lower rating for the PAB than the standard index. This is also the case for the CC Index, although the degrees of low-score overweighting and high-score underweighting are not as pronounced as those of the PAB. The observed CWDs using REF and BB constituent ratings, to some extent, explain the puzzling result of the negative coefficient estimate of GCE for PAB (see columns 2, 3, 6, Table 3).

The statistics of ENV ratings (Appendix B, Panel A) confirm the finding of this analysis: the REF and BB ENV scores for the PAB have lower average values than those of the standard index. However, it should be noted that the strict criteria of both the SRI and SRI PAB are strong enough for them to have a significantly higher ENV score than the standard index.[39]

Panel A: CWD between green and standard indices for ENV rating-ranked constituents in January, 2015



Panel B: CWD between green and standard indices for ENV rating-ranked constituents in July, 2019

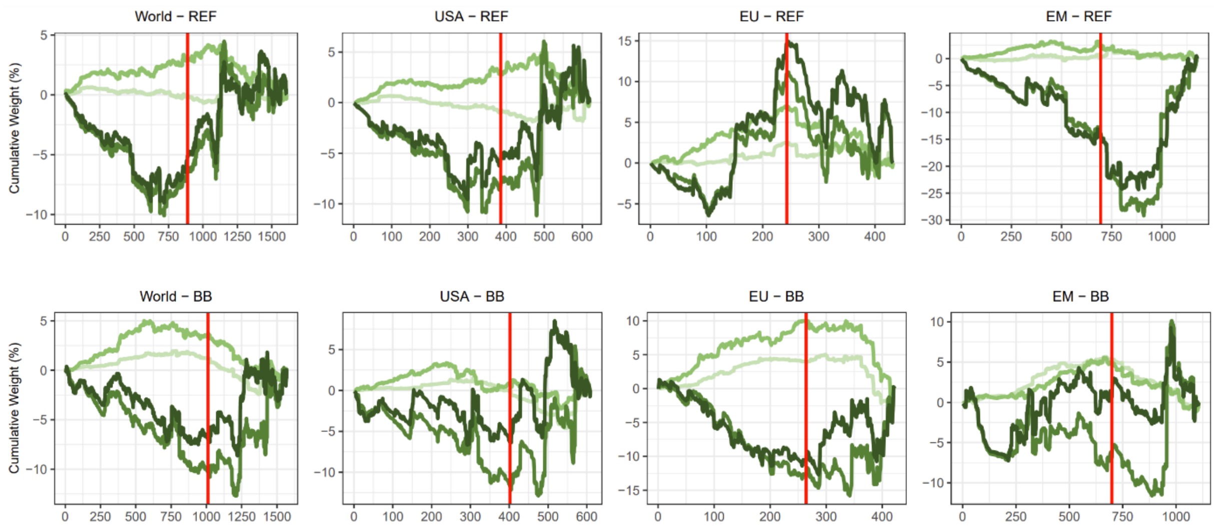
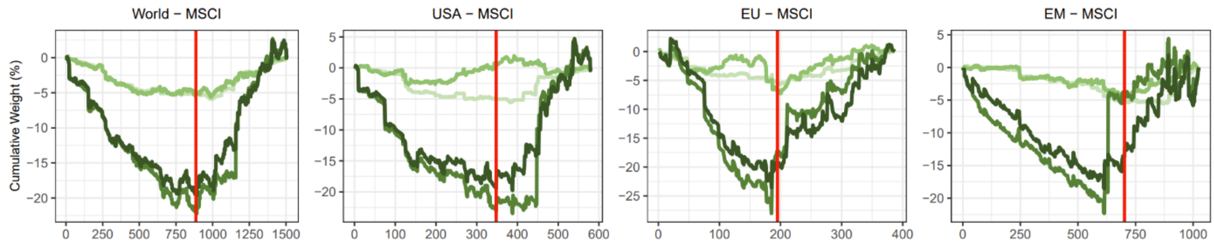


Figure G.1 Cumulative Weighting Difference (CWD) between green and standard indices for ENV rating-ranked constituents

[39] This study performs an alternative CWD analysis of ESG scores assigned by REF, BB, and MSCI at identical time points and finds qualitatively similar results.



Panel C: CWD between green and standard indices for ENV rating-ranked constituents in July, 2022

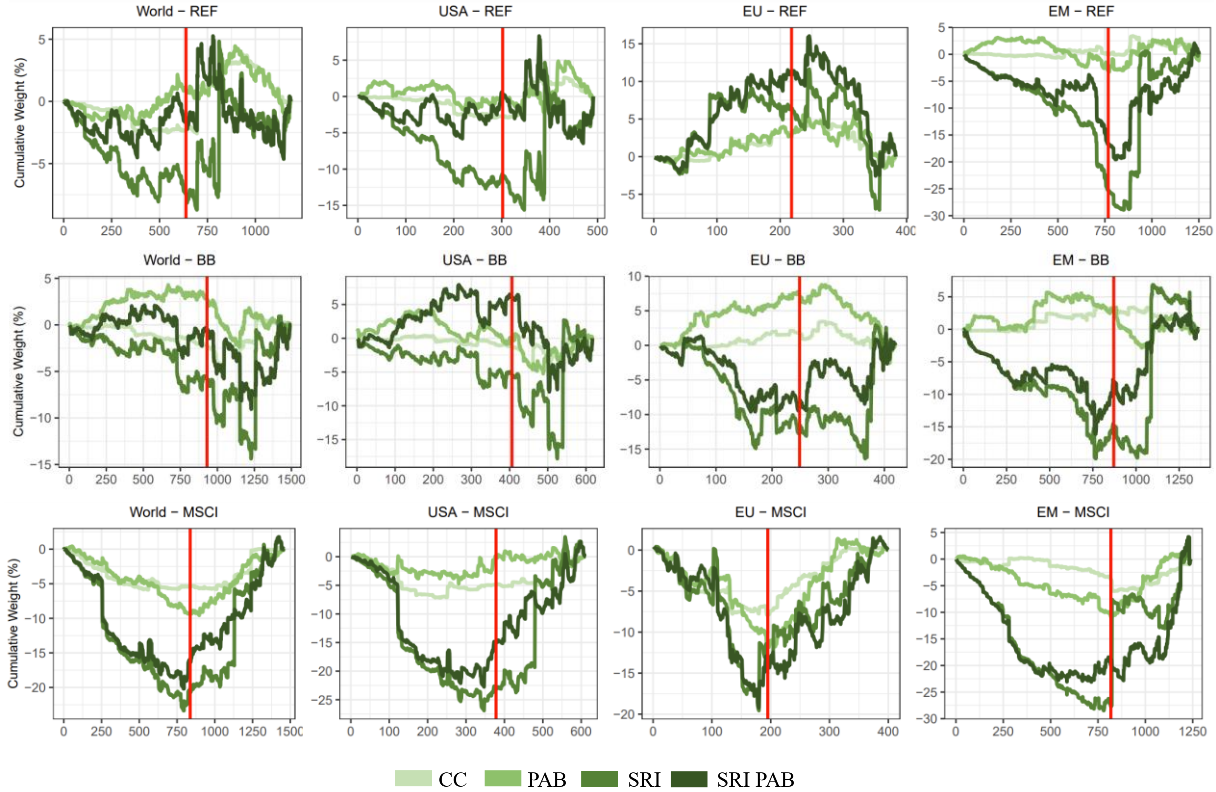


Figure G.1 Cumulative Weighting Difference (CWD) between green and standard indices for ENV rating-ranked constituents (continued)

This figure features the Cumulative Weighting Difference (CWD) of the Environmental (ENV) Scores reported by Refinitiv (REF), Bloomberg (BB) and MSCI, across four regions (the World, the U.S., Europe and Emerging Markets). We conduct this analysis at three points of time: January 2015 - the beginning of our study period (Panel A), July 2019 - the middle of our study period and six months before the spread of the COVID-19 pandemic (Panel B), and July 2022 - shortly after the Ukraine war started (Panel C).

Appendix H. Environmental Rating Performance Attribution

H.1 Performance Attribution Method

In our analysis of the environmental (ENV) ratings of MSCI green indices, we utilize constituents' weights provided by MSCI and constituents' rating scores issued by two providers, REF and BB. It should be emphasised that MSCI ratings are not used in OLS regressions to avoid concerns of endogeneity. The OLS analysis of the effect of MSCI's GCEs on an index's ENV (alternatively, ESG) ratings issued by REF and BB (equation 3) generates a puzzling result, which raises a natural question regarding the sources of rating differences across providers. Specifically, is a green index's ENV excess rating driven by departures from its respective parent index's sector allocation, and/or departures from the parent index within each sector (constituent selection)?

A green index's higher ENV rating (excess rating) can be achieved by underweighting (overweighting) carbon-intensive (green) sectors (relative to the parent index) and/or investing in green stocks at the "right" time. For example, MSCI may have higher (lower) weightings (compared with the standard parent index) in sectors they expect to environmentally perform well (poorly). Alternatively, MSCI may exhibit skill in selecting outstanding (greenest) constituents from the best environmentally performing sector.

Following Bodie, Kane and Marcus (2014, p. 865)'s approach to decompose a portfolio's excess return and similar to Jacob and Wilkens (2021) attribution analysis, we decompose the ENV excess rating at time *t* of a green index g*i* relative to its parent index *pi* into a sum of two components namely sector contribution and stock (constituent) selection:

$$Excess\ rating_{gi,pi,t} = gi_{pr} - pi_{pr}$$

$$Excess\ rating_{gi,pi,t} = \textstyle\sum_{s=1}^{11}(\omega_{gi,s,t} - \omega_{pi,s,t})pr_{pi,s,t} + \sum_{s=1}^{11}\omega_{gi,s,t}(pr_{gi,s,t} - pr_{pi,s,t})$$

$$gi_{pr} - pi_{pr} = \sum_{s=1}^{11}(\omega_{gi,s,t} - \omega_{pi,s,t})pr_{pi,s,t} + \sum_{s=1}^{11}\omega_{gi,s,t}(pr_{gi,s,t} - pr_{pi,s,t}) \qquad (5a)$$

The first term on the right-hand side of equation (5b) measures the effect of sector allocation across the 11 sectors in this study. This term can be estimated directly from the weight ω of each sector *s* in a green index *gi* ($\omega_{gi,s,t}$), the weight of sector *s* in the parent index *pi* ($\omega_{pi,s,t}$), and the provider rating *pr* of the parent index for sector *s*, $pr_{pi,s,t}$.

$$Sector\ allocation = \sum_{s=1}^{11}(\omega_{gi,s,t} - \omega_{pi,s,t})pr_{pi,s,t} \qquad (5b)$$

The second term on the right-hand side of equation (5c) captures the stock (constituent) selection, and can be derived based on the firms-weighted rating of sector s in the green index gi ($pr_{gi,s,t}$), the firms-weighted provider rating pr of sector s in the parent index pi ($pr_{pi,s,t}$), and the weight of sector s in the green index gi ($\omega_{gi,s,t}$).

$$Stock\ selection = \sum_{s=1}^{11} \omega_{gi,s,t}\left(pr_{gi,s,t} - pr_{pi,s,t}\right) \qquad (5c)$$

H.2 Environmental Rating Performance Attribution Results

The CWD analysis (Appendix G) leads us to explore the drivers of the excess rating between a green index and the standard index, and how the contributions of these drivers vary across the three providers (Refinitiv, Bloomberg, MSCI). Each green index's excess rating (estimated as in equation 5a) is decomposed into the sum of two components: sector allocation (equation 5b), and constituent (stock) selection (equation 5c). Table H.1 summarises the results of the ENV rating performance attribution analysis for the green indices between 2015 and 2022 - the last year when the ENV rating data is available.

While there are some variations in the contributions of each component across providers and across years, Table H.1 features several striking features. For MSCI, all four green indices demonstrate positive excess ratings varying between 2.66 and 13.14 units (relative to the standard index). Furthermore, MSCI consistently does well in delivering positive values for both components (except the sector allocation of SRI). This confirms that MSCI's decisions are consistent with their published index construction criteria. For REF and BB, the two restrictive indices SRI and SRI PAB consistently generate positive excess ratings. In contrast, we observe negative sector allocation values and negative excess ratings for the CC and PAB in all years. Notably, PAB is the only index that creates negative values for both constituent (stock) selection and sector allocation and delivers negative ratings across years. On average, CC performed slightly better than PAB as CC's stock selection mean value is positive, and the degree of rating underperformance is less pronounced. This analysis provides further insights into the puzzling negative effect of MSCI's GCE on the REF rating for PAB (Table 3).

Table H.1 Attribution analysis of environmental (ENV) rating differential between a green index and its standard index, January 2015-December 2022

	Constituents (Stock) Selection (Equation 5c)				Sector Allocation (Equation 5b)				ENV excess rating (Equation 5a)			
	CC (1)	PAB (2)	SRI (3)	SRI PAB (4)	CC (5)	PAB (6)	SRI (7)	SRI PAB (8)	CC (9)	PAB (10)	SRI (11)	SRI PAB (12)
Year						**Panel A:** Refinitiv ENV ratings						
2015	-0.05	-3.64	7.22	6.75	-0.43	-0.44	-0.28	-0.2	-0.48	-4.08	6.94	6.55
2016	0.02	-2.29	9.23	7.63	-0.66	-0.88	-0.51	0.08	-0.64	-3.17	8.72	7.71
2017	-0.11	-2.29	8.8	7.36	-0.51	-0.97	-0.06	0.47	-0.62	-3.26	8.74	7.83
2018	-0.15	-2.05	8.93	7.47	-1.03	-1.17	-0.05	-0.32	-1.18	-3.22	8.88	7.15
2019	0.67	-1.65	6.02	5.09	-0.87	-0.6	0.62	0.53	-0.2	-2.25	6.64	5.62
2020	0.83	-0.65	3.51	1.39	-0.7	-0.67	0.59	0.58	0.13	-1.32	4.1	1.97
2021	0.14	-0.26	3.1	0.47	-0.57	-0.85	0.19	0.69	-0.43	-1.11	3.29	1.16
2022	0.21	-0.17	3.66	0.88	-0.45	-0.92	0.18	0.85	-0.24	-1.09	3.84	1.73
Average	0.20	-1.63	6.31	4.63	-0.65	-0.81	0.09	0.34	-0.46	-2.44	6.39	4.97
Year						**Panel B:** Bloomberg ENV ratings						
2015	0.72	-0.55	9.34	14.63	-2.3	-3.82	0.39	-1.87	-1.58	-4.37	9.73	12.76
2016	0.99	-0.71	12.35	14.56	-2.9	-4.81	-0.03	-4.67	-1.91	-5.52	12.32	9.89
2017	0.56	-1.51	12.46	12.58	-3.72	-4.78	0.72	-5.83	-3.16	-6.29	13.18	6.75
2018	0.16	-0.89	16.35	15.58	-4.05	-4.99	0.58	-6.16	-3.89	-5.88	16.93	9.42
2019	0.49	-1.14	10.02	10.32	-2.61	-4.16	1.64	-4.48	-2.12	-5.3	11.66	5.84
2020	0.03	-1.08	10.57	9.63	-1.3	-2.31	0.29	-4.51	-1.27	-3.39	10.86	5.12
2021	0.25	-1.67	7.24	5.26	-0.69	-1.4	-0.45	-3.87	-0.44	-3.07	6.79	1.39
2022	0.33	-1.74	7.60	6.24	-0.81	-1.37	-0.35	-3.81	-0.48	-3.11	7.25	2.43
Average	0.44	-1.16	10.74	11.10	-2.30	-3.46	0.35	-4.40	-1.86	-4.62	11.09	6.70
Year						**Panel C:** MSCI ENV ratings						
2015	1.55	0.72	12.13	12.71	2.92	2.13	1.11	2.68	4.47	2.85	13.24	15.39
2016	1.64	0.41	14.68	13.3	2.43	1.94	0.45	2.09	4.07	2.35	15.13	15.39
2017	1.39	0.3	13.58	12.37	2.35	1.92	-0.56	1.97	3.74	2.22	13.02	14.34
2018	1.08	0.44	12.8	11.16	2.45	1.8	-0.53	1.53	3.53	2.24	12.27	12.69
2019	1.66	0.9	14.17	12.77	2.04	1.29	-1.78	-0.53	3.7	2.19	12.39	12.24
2020	1.88	1.77	12.28	12.16	1.01	0.73	-2.23	-1.01	2.89	2.5	10.05	11.15
2021	2.5	2.41	12.49	11.37	0.77	1.04	-1.18	0.01	3.27	3.45	11.31	11.38
2022	2.92	2.64	12.44	11.46	1.14	0.8	0.12	1.09	4.06	3.44	12.56	12.55
Average	1.83	1.20	13.07	12.16	1.89	1.46	-0.58	0.98	3.72	2.66	12.50	13.14

This table shows the results of the environmental rating attribution analysis which determines the sources of the difference (excess rating) between the environmental (ENV) rating of a green index and that of the MSCI Standard Index. This index-level analysis is conducted using constituents' ENV ratings (available to 2022) provided by Refinitiv (Panel A), Bloomberg (Panel B) and MSCI (Panel C). The rating differential (excess rating) between a green index and the standard index is attributed to two components, namely constituents (stock) selection and sector allocation. Columns 1-4 provide the values of the stock selection component for four green indices, each is estimated as in Equation (5c). This component indicates whether constituents (stocks) of a green index are "greener" (have greater ENV ratings) than the constituents (stocks) of the MSCI Standard Index. Columns 5-8 show the values of the sector allocation component for four green indices, each is estimated as in Equation (5b). This component shows whether green (standard) sectors are overweighted (underweighted) in a green index (compared to the MSCI Standard Index). Columns 9-12 show the differential between the ENV rating *pr* of a green index g*i* and that of the MSCI Standard Index *pi* (excess rating, Equation 5a). The rating differential (excess rating) is equal to the total value of the stock selection and sector allocation components.

Appendix I. Variable definitions and data sources

Emissions Data		
Variable	**Descriptions/ Definitions sourced from data providers**	**Source**
	Refinitiv	
Total Emissions (Scope 1+2)	Total Carbon dioxide (CO_2) and CO_2 equivalents emission in tonnes. Total sum of Scope 1 and Scope 2 emissions.	ENERDP023
Direct Emissions (Scope 1)	Direct emissions of CO_2 and CO_2 equivalents emission in tonnes from sources that are owned or controlled by the company.	ENERDP024
Indirect Emissions (Scope 2)	Indirect emissions of CO_2 and CO_2 equivalents emission in tonnes from consumption of purchased electricity, heat or steam which occur at the facility where electricity, steam or heat is generated.	ENERDP025
CO_2 Intensity (Scope 1+2)	Total CO_2 and CO_2 equivalents emission in tonnes divided by net sales or revenue in US dollars in millions.	ENERO03V
	Bloomberg	
Direct Emissions (Scope 1)	Amount of scope 1 greenhouse gas (GHG) emissions of the company, in thousands of metric tonnes of carbon dioxide equivalent (CO_2). Greenhouse gas emissions are defined as those gases which contribute to the trapping of heat in the Earth's atmosphere, including carbon dioxide (CO_2), methane, nitrous oxide, and others. Emissions reported as CO_2 only will NOT be captured in this field. Examples of scope 1 emissions include emissions from combustion in owned or controlled boilers, furnaces, vehicles, and emissions from production in owned or controlled process equipment.	GHG_SCOPE_1
Indirect Emissions (Scope 2)	Amount of scope 2 greenhouse gas (GHG) emissions of the company, in thousands of metric tonnes of carbon dioxide equivalent (CO_2), using the market-based accounting method. Greenhouse gases are defined as those gases which contribute to the trapping of heat in the Earth's atmosphere, including carbon dioxide (CO_2), methane, nitrous oxide, and others. Emissions reported as CO_2 only will NOT be captured in this field. Scope 2 emissions are those emitted as a consequence of the activities of the reporting entity, but that occur at sources owned or controlled by another entity. The main sources of scope 2 emissions are emissions from purchased electricity, steam and/or heating/cooling. The market-based method for scope 2 GHG derives emissions factors from contractual instruments.	GHG_SCOPE_2_ MARKET_BASED
CO_2 Intensity (Scope 1)	Greenhouse gas (GHG) intensity calculated as metric tonnes of greenhouse gases in carbon dioxide equivalent (CO_2) emitted from direct operations per million of sales revenue in the company's reporting currency. Scope 1 Emissions are those emitted from sources that are owned or controlled by the reporting entity. Examples of Direct Emissions include emissions from combustion in owned or controlled boilers, furnaces, vehicles, emissions from chemical production in owned or controlled process equipment. To compare companies globally, this ratio should be converted to a common currency. Ratio is calculated based on data items disclosed in company filings. Calculated as: GHG Scope 1 Emissions*1000 / Sales	GHG_SCOPE_1_ INTENSITY_ PER_SALES
CO_2 Intensity (Scope 2)	Greenhouse gas (GHG) intensity calculated as metric tonnes of greenhouse gases in carbon dioxide equivalent (CO_2) emitted from indirect operations per million of sales revenue in the company's reporting currency. Scope 2 Emissions are those emitted that are a consequence of the activities of the reporting entity, but occur at sources owned or controlled by another entity. The main source of Indirect Emissions is emissions from purchased electricity, steam and/or heating/cooling. These emissions physically occur at the facility where electricity/steam/heating/cooling is generated. To compare companies globally, this ratio should be converted to a common currency. Ratio is calculated based on data items disclosed in company filings. Calculated as: GHG Scope 2 Emissions*1000 / Sales	GHG_SCOPE_2_ INTENSITY_ PER_SALES

Appendix I. Variable definitions and data sources (continued)

Emissions Data		
Variable	**Descriptions/ Definitions sourced from data providers**	**Source**
	Carbon Disclosure Project	
Direct Emissions (Scope 1)	Total global amount of scope 1 emissions emitted by the company, measured in millions of metric tons of carbon dioxide equivalent (mt CO_2). Scope 1 emissions are direct GHG (greenhouse gas) emissions from sources that are owned or operated by the company. Sources include combustion facilities, company owned or operated transportation, and physical or chemical processes. The information is directly from the company's response to the CDP climate change information request.	CDP_SCOPE_1_ EMISSIONS_ GLOBALLY
Indirect Emissions (Scope 2)	Gross global market-based Scope 2 emissions in millions of metric tonnes of carbon dioxide equivalent (CO_2). Market based Scope 2 uses supplier specific emission factor(s) in calculation, e.g. through a contractual instrument. The information is directly from the company's response to the CDP climate change information request.	CDP_MKT_ BASED_SCOPE_2_ EMISSIONS
CO_2 Intensity (Scope 1+2)	Revenue emission intensity figure for the reporting period (gross Scope 1+2 in metric tonnes of CO_2-equivalent / unit currency total revenue). The information is directly from the company's response to the CDP climate change information request.	CDP_REVENUE_ EMISSION_ INTENSITY

ESG Data		
	Refinitiv (A higher rating reflects a better environmental performance assessment)	
Environmental Score	Refinitiv's Environment Pillar Score is the weighted average relative rating of a company based on the reported environmental information and the resulting three environmental category scores (Emissions, Resource Use, Environmental Innovation).	ENSCORE
ESG Score	Refinitiv's ESG Score is an overall company score based on the self-reported information in the environmental, social and corporate governance pillars.	TRESGS
Controversies Score	ESG controversies category score measures a company's exposure to environmental, social and governance controversies and negative events reflected in global media.	TRESGCCS
	Bloomberg (A higher rating reflects a better environmental performance assessment)	
Environmental Score	Proprietary Bloomberg score from 0 to 10 to assess the company's aggregated environmental per-formance.	ENVIRONMENTAL_ SCORE
ESG Score	Provisional proprietary Bloomberg score from 0 to 10 for the assessment of the company's aggregated ESG performance (Environmental, Social & Governance, ESG).	ESG_SCORE
	MSCI (A higher rating reflects a better environmental performance assessment)	
Environmental Score	The score represents the weighted average of the scores received on all the Key Issues contributing to the final rating of the company.	WEIGHTED_ AVERAGE_SCORE
ESG Score	The Environmental Pillar Score represents the weighted average of all Key Issues that fall under the Environment Pillar.	ENVIRONMENTAL_ PILLAR_SCORE

Appendix I. Variable definitions and data sources (continued)

Firm-level Data (Refinitiv)		
Variable	**Descriptions**	**Source**
Market Value	The Market Value of a security calculated by multiplying the number of shares (NOSH) and security price (P).	MV

Control Variables		
Variable	**Descriptions**	**Source**
Sector composite	Constituents in the sample are from 11 sectors: Health Care, Consumer Discretionary, Materials, Industrials, Information Technology, Financials, Consumer Staples, Energy, Utilities, Communication Services, and Real Estate. The sector composite value is the sector monthly weights in each index, measured based on the firms' monthly weights. If 15% of the firms in an index are Financials, the value of the Financial sector in that index is 15%. As constituents (firms)' weights change each month and constituents in an index also change throughout the year, an index's sector composite value changes each month.	MSCI Index Dataset
Human Development Index (HDI) composite	In each year, we categorise countries in our sample into 5 groups based on their numeric HDI values (<=0.6, 0.6-0.7, 0.7-0.8, 0.8-0.9, ≥0.9). The reference group is ≥0.9. The HDI value of firm *j* at time *t* is equal to the HDI value of the country where firm *j* is domiciled at time *t*. If HDI value is not available for a year, we take the most recent year's value. The index's value for an HDI group is the sum of firms weights in that HDI group. If 15% of the firms in an index have a HDI value between 0.7-0.8, the value of the HDI group (0.7-0.8) is 15%. As constituents (firms)' weights change each month and constituents in an index also change throughout the year, an index's HDI value changes each month. We gather HDI data from the United Nations Development Program (UNDP).	MSCI Index Dataset, UNDP
Post Launch dummy	A dummy variable indicating whether an index has been officially launched prior to the analysed month. A value of 1 indicates the index has been launched, while 0 indicates it has not been launched yet.	MSCI
Covid dummy	A dummy variable that captures the impact of the COVID-19 pandemic, defined from February 2020 to January 2022, to reflect the period of the global outbreak.	WHO (2020)
Ukraine dummy	A dummy variable which marks the outbreak of the war in Ukraine, starting from the end of February 2022 and continuing until the last observation in the dataset.	Self-created

Paper 3 – Small Deviations, Large Reductions: An Optimization Approach to Low-Tracking Error Decarbonized European Equity Portfolios

Abstract[*]

This paper examines the extent to which European equity portfolios can be decarbonized while maintaining a low tracking error. Using reported Refinitiv emissions and a transparent multi-factor risk model, I construct optimized low-carbon portfolios of the STOXX Europe 600 over 2011–2024. Portfolios targeting Scope 1–2 carbon intensity achieve reductions of 61%, 84%, and 93% for tracking-error budgets of 0.5%, 1.0%, and 2.0%, respectively, with realized tracking errors close to target. I observe no performance penalty; Sharpe and information ratios are slightly higher than the benchmark. Between 67 and 85% of carbon reductions and most active returns come from within-sector stock selection rather than sector tilts. Results are robust to alternative carbon metrics, scopes, sample periods, and reporting currencies. Substantial decarbonization of European equity portfolios is feasible without sacrificing diversification or performance.

Keywords: Low-carbon investing; Portfolio optimization; Tracking error; Carbon intensity; Scope 1–2 emissions; Climate risk

[*] This paper is single-authored. It is currently under consideration at the *Journal of Sustainable Finance & Investment*.

1 Introduction

Global investor demand for climate-aligned products has grown rapidly with assets in public climate-themed funds having increased 10-fold over the last six years, from about \$60 billion in 2019 to \$625 billion as of October 2025 (Lee et al., 2025). One type of climate-themed fund is a low-carbon strategy, which promises to reduce the carbon emissions of a portfolio relative to its benchmark. Investors have multiple, and often overlapping, motivations to allocate to these low-carbon portfolios or similar ESG products. Prominent motivations for retail investors include ethical considerations as well as prosocial and green preferences (Giglio et al., 2025; Siemroth and Hornuf, 2023; Gutsche, Wetzel and Ziegler, 2023; Heeb et al., 2023; Riedl and Smeets, 2017). Likewise, institutional investors increasingly integrate climate considerations into their investments, motivated by ethical and financial concerns (Robeco, 2025; Kolasa and Sautner, 2025; Martini, 2021; Krueger, Sautner and Starks, 2020), but also by pressure, including from regulators, to integrate climate risks and decarbonization objectives into their portfolios (Robeco, 2025; Otchere et al., 2025; Balp and Strampelli, 2022; Ameli et al., 2020) or for reputational and competitive motives (Cao et al., 2025; Mazzolini et al., 2025; Alda et al., 2022; Krueger, Sautner and Starks, 2020). Both groups, however, are also interested in reducing their exposure to climate-related risks, including transition risks, physical climate risks, or litigation risks (Giglio et al., 2025; Morgan Stanley Institute for Sustainable Investing, 2025; Setzer and Higham, 2025; Bell and Taylor, 2024; FTSE Russell, 2022).

Low-carbon portfolios can be constructed with various methods that all start from a chosen benchmark and then reduce a carbon metric such as carbon intensity (CI) or carbon footprint (CF), typically based on Scope 1–2 (S1–2) or Scope 1–3 (S1–3) emissions. A common goal is to minimize deviation from the benchmark, which is usually measured by the tracking error (TE). One basic construction method is screening, in which firms whose emissions exceed fixed thresholds are removed and the remaining index constituents are then re-weighted. Bouchey et al. (2024), for example, screen the S&P Developed Markets index on S1–3 emissions (using their “heavy” screen with thresholds of 500 for S1–2 and 300 for S1–3) and obtain a CI reduction of 58% with a realized (ex post) TE close to 1%. Simple fossil-fuel divestment screens on broad indices, such as the S&P 500 analyzed by Sireklove (2016), also generate only modest TEs (1.4–1.9%) and small long-run performance differences (3 to 11 bps), although the author does not report portfolio-level carbon reductions.

A second group is exclusion and best-in-class strategies. Exclusion strategies remove a share of companies with the highest carbon metric, while a best-in-class strategy overweights the companies with the lowest carbon metrics. A common approach is to combine these two strategies, as in Jondeau et al. (2025), who reinvest the excluded proportion solely into the companies with the lowest metrics within each sector and region. Jondeau et al. (2025) show that excluding the most carbon-intensive Scope 1–3 companies and reinvesting within sectors/regions can lower CI by roughly 60–75% relative to MSCI ACWI (and its Europe subset), with annual TEs between about 1% and 3% depending on the exclusion threshold. Fahlenbrach and Jondeau (2023) apply a related exclusion approach to the Swiss National Bank's (SNB) U.S. equity portfolio (Scope 1–2), achieving about a 61% reduction in CI and a 53% reduction in CF at a very low TE (0.23% versus the SNB portfolio and 1.19% versus MSCI USA). Andersson et al. (2016) analyze the MSCI Europe Low Carbon Index, obtaining Scope 1–2 carbon-intensity reductions of 52–62% versus the MSCI Europe Index for a low TE of 0.7%.

Optimization-based approaches treat carbon metrics as an explicit objective or constraint in a tracking-error-minimization problem. Using multi-factor risk models, these studies minimize carbon intensity or carbon footprint subject to ex-ante TE budgets and standard investability constraints. Bouchey et al. (2024) minimize Scope 1–3 carbon intensity of the S&P Developed Markets under TE constraints between 0.25% and 1%, achieving CI reductions of about 30–60%. Loosening sector and industry bounds pushes the reduction up to roughly two-thirds at TEs around 1–1.5%. Bender et al. (2019) optimize MSCI World portfolios on Scope 1–3 metrics and report CI reductions of roughly 64–73% for ex-ante TE targets of 1%, 2%, and 3%, with realized TE of 1.4–3.2%. Andersson et al. (2016) construct an optimized MSCI Europe portfolio that reaches an 82% reduction in Scope 1–2 CI for an ex-post TE of about 0.9%. Blitz et al. (2024) use a 3D optimization that jointly optimizes expected return, risk, and sustainability for MSCI World portfolios, showing that including Scope 1–2 carbon footprint, measured as emissions divided by enterprise value including cash (EVIC), in the objective function can deliver carbon footprint reductions of 59–73% at tracking-error targets of 0.5–1.0%. Online Appendix A summarizes the carbon reductions and tracking errors reported in these studies. Beyond these broad-market applications, related optimization-based work extends this logic to minimum-volatility and net-zero benchmarks: Bender et al. (2020) decarbonize MSCI World minimum-volatility portfolios and obtain carbon-intensity cuts close to 90% while preserving their lower total risk. Bolton et al. (2022) and Kolle et al. (2022) design Paris-aligned MSCI World/Europe portfolios that reduce carbon intensity by around 50–90% for ex-ante tracking errors on the order of 1–2%.

Beyond the equity index strategies summarized in Online Appendix A, a broader body of literature studies related ways to decarbonize portfolios and to price transition risk. Cenedese, Han, and Kacperczyk (2023) construct forward-looking “distance-to-exit” net-zero portfolios that cut carbon footprints by 40–95% while remaining diversified, and further show that “distance-to-exit” is priced as a transition-risk factor. In et al. (2019) form efficient-minus-inefficient long–short portfolios sorted on firm-level carbon intensity (Scope 1–3) and find abnormal returns for carbon-efficient firms. Cheng, Jondeau, and Mojon (2022) extend net-zero portfolio construction to sovereign bond benchmarks, achieving around 40% reductions in sovereign carbon intensity at TEs of roughly 1%. At the multi-asset level, Jia et al. (2025) illustrate how reallocating from standard indices into low-carbon and Paris-aligned variants across asset classes can halve portfolio-level Scope 1–2 CI. While these studies confirm that substantial emissions reductions are achievable across asset classes, they do not address the design of low-tracking-error equity portfolios for European markets, which is the goal of this study.

Despite the extensive literature, several gaps remain for investors interested in low-carbon European equity portfolios. First, most portfolio or index-style decarbonization studies focus on global benchmarks such as the MSCI ACWI, MSCI World or S&P Developed Markets. Europe only appears as one region within a global portfolio (Jondeau et al., 2025; Bender et al., 2019) or in early studies on MSCI Europe (Andersson et al., 2016). Second, most studies focus on Scope 1–3 metrics and rarely separate Scope 1–2 from Scope 3, even though Scope 3 emissions are typically estimated rather than reported and are subject to much greater measurement error and provider disagreement than Scope 1–2 data (Popescu et al., 2023; Busch et al., 2022). Third, the construction of low-carbon indices in the literature almost always relies on carbon datasets that include estimated emissions, such as S&P Global Trucost (Jondeau et al., 2025; Bouchey et al., 2024; Fahlenbrach and Jondeau, 2023; Bender et al., 2020; Bender et al., 2019; Andersson et al., 2016). Fourth, many studies use commercial multifactor risk models and optimizers such as MSCI Barra (Bouchey et al., 2024; Andersson et al., 2016) or Axioma (Bender et al., 2020; Bender et al., 2019) to estimate ex-ante tracking error and construct portfolios. This makes their implementation difficult to replicate for investors who do not have access to these tools.

This study contributes to the literature by filling the following gaps. First, I focus explicitly on European equity universes by constructing and evaluating low-carbon variants of the STOXX Europe 600 and, as a robustness check, of the STOXX Europe Total Market Index (TMI) over 2011–2024, starting when sufficient Scope 1–2 reporting coverage became available. Second, I separately analyze and optimize portfolio-level emissions for Scope 1–2 and Scope 1–3, and report carbon reductions separately for both scopes. Third, I rely only on reported emissions and revenues from Refinitiv (LSEG) for all carbon metrics. Fourth, I construct low-carbon portfolios using a self-built optimizer based on free-float value-weighted parent indices, return covariance estimates, and explicit tracking-error constraints, without relying on commercial risk models such as Axioma or MSCI Barra. With this approach, I aim to answer the following research question:

How much can portfolio-level carbon emissions be reduced in a European equity universe while maintaining a low tracking error relative to a free-float value-weighted parent index?

The main finding is that European equity portfolios can achieve Scope 1–2 carbon-intensity reductions of 60 to over 90 percent at realized tracking errors of 0.6 to 2.1 percent per year, with no evidence of a performance penalty. Across three scenarios (Conservative, Balanced, and Aggressive, defined by progressively looser ex-ante tracking-error targets of 0.5%, 1%, and 2%), information ratios range from approximately 0.5 to 0.8, and total volatility remains virtually identical to the benchmark. These results are robust to alternative carbon objectives (intensity versus footprint), emissions scopes, parent indices, time windows, and reporting currencies. Broader sustainability metrics, including Refinitiv and MSCI ESG scores, are maintained or modestly improved.

These findings have direct implications for practitioners. A decomposition of the carbon reductions reveals that 67 to 85 percent of the total reduction stems from within-sector stock selection rather than sector tilts. A Brinson-style performance attribution delivers a similar conclusion: within-sector selection explains most of the active return, while sector-allocation effects remain small. This implies that asset managers can maintain sector exposures close to standard benchmarks and implement a benchmark-aware, portfolio optimization-based best-in-class tilt toward cleaner stocks within each sector, an approach that is consistent with many existing investment processes. Because the framework relies only on reported emissions and a transparent optimization approach, it can be implemented without proprietary risk models or commercial climate-index products, lowering the barrier for smaller asset managers and asset owners who seek to decarbonize their portfolios.

The remainder of the paper proceeds as follows. Section 2 describes the data, including the construction of the investment universe and carbon metrics. Section 3 presents the optimization framework, constraint structure, and scenario design. Section 4 reports the main results and robustness tests. Section 5 concludes.

2 Data

2.1 Universes and sample periods

My sample universe consists of European equities which are represented by the STOXX Europe 600 and STOXX Europe Total Market Index (TMI). The STOXX Europe 600 contains large-, mid-, and some small-cap companies from 17 developed European countries and covers about 90% of the free-float market capitalization of the developed European equity market. The STOXX Europe TMI covers approximately 95% of the free-float market value (STOXX, 2024). For both indices, I obtain the historical monthly constituent lists and the corresponding free-float market value (MVFF) directly from Refinitiv starting in January 2005.[1] At each month-end t, I identify all stocks that belong to the STOXX Europe 600 or STOXX Europe TMI and construct index weights as

$$\omega_{i,t} = \frac{MVFF_{i,t}}{\sum_{j=1}^{N_t} MVFF_{j,t}} \tag{1}$$

where $\omega_{i,t}$ is the free-float weight of stock i at time t, $MVFF_{i,t}$ is the free-float market value in euros, and N_t is the number of constituents in the respective index at time t. Thus, $\omega_{i,t}$ measures the share of stock i's free-float market value relative to the total free-float market value of all index members at that month-end.[2] Because I use the historical constituent lists and MVFF as of each date, the reconstructed indices are free of survivorship bias. All prices and returns are expressed in euros. Local currency data were converted using end-of-day exchange rates from Refinitiv. In a robustness test, I recompute the main results in US dollars. Additionally, the currency choice does not affect the carbon reduction because any currency conversion cancels out in both the portfolio and the benchmark.

The empirical analysis focuses on the period from January 2011 to December 2024. At the start of each calendar year, I form optimized low-carbon portfolios based on the previous year-end values and hold them from January to December; portfolios are therefore rebalanced annually (see Section 3 for details). The STOXX Europe 600 contains between 590 and 600 constituents, with a median stock weight of 0.07%, while the STOXX Europe TMI grew from 961 constituents in January 2011 to 1,882 by the end of 2024 and has a median weight of 0.02%, reflecting its broader small-cap coverage. Sector and country weight distributions averaged over

[1] Approximately 0.3% of the index weights lack an MVFF value. For these observations, I impute MVFF using sector-specific ratios of market value to free-float market value.

[2] As a robustness check, I replicate the STOXX Europe 600 from 2005 to 2024 using these weights. The replicated index's average monthly return and end-of-period value (0.65% and 2.83) are virtually identical to the official index.

2011–2024 are reported in Appendix Table A.1. Both indices are dominated by Financials, Health Care, and Industrials, and by firms from the United Kingdom, France, Switzerland, and Germany.

2.2 Carbon data

2.2.1 Scopes and portfolio metrics

I obtain Scope 1 (direct), Scope 2 (indirect, purchased electricity and heat) and Scope 3 (indirect, value-chain) greenhouse gas (GHG) emissions for all stocks that have ever belonged to the STOXX Europe 600 or STOXX Europe TMI from Refinitiv.[3] As firms typically report their GHG emissions and their revenue with a lag, I lag both variables by one year to mimic the information set of an investor and avoid look-ahead bias. Therefore, for portfolios held in year t, I use emissions and revenue reported for year $t–1$. For each stock i at time t, $E_{i,t-1}^{CS}$ denotes the absolute emissions E for a composite scope (S1–2, S1–3), measured in tonnes of CO_2-equivalents. Using $E_{i,t-1}^{CS}$, firm revenue $Rev_{i,t-1}$ and market value $MV_{i,t}$, I construct two firm-level carbon metrics:

$$CI_{i,t}^{cs} = \frac{E_{i,t-1}^{cs}}{Rev_{i,t-1}} \tag{2}$$

$$CF_{i,t}^{cs} = \frac{E_{i,t-1}^{cs}}{MV_{i,t}} \tag{3}$$

Here, $CI_{i,t}^{cs}$ is revenue-based carbon intensity and $CF_{i,t}^{cs}$ is market-value-based carbon footprint. For example, if a firm has a $CI_{i,t}^{S1-2}$ of 150, this means it emits on average 150 tonnes of CO_2-equivalent Scope 1 and 2 emissions for every 1 million EUR of revenue. Naturally, a firm in a carbon-intensive sector such as Utilities will tend to have high Scope 1 emissions, while a technology firm may have relatively low Scope 1 but perhaps higher Scope 2 emissions to power its data centers.

In the main analysis, I focus on Scope 1–2, which are directly controlled, widely reported, and form the basis of most portfolio-level climate metrics recommended by the Task Force on Climate-related Financial Disclosures (TCFD, 2021).[4] Scope 3 emissions, in contrast, are often estimated and subject to greater measurement error and provider disagreement (Popescu et al., 2023; Busch et al., 2022). However, as a robustness check, I additionally investigate Scope 1–

[3] See Appendix E for the scope's definitions and Datastream Codes.
[4] TCFD (2021) encourages disclosure of Scope 3 emissions where they are relevant and sufficiently reliable.

3 emission reductions, both (i) as a byproduct when reporting the results of the Scope 1–2 optimization and (ii) as the direct objective of a separate Scope 1–3 optimization.

To analyze the portfolio exposure of these metrics at index level, I aggregate $\mathrm{CI}_{i,t}^{cs}$ and $\mathrm{CF}_{i,t}^{cs}$ using the free-float value weights $\omega_{i,t}$ from equation (1):

$$CI_{I,t}^{cs} = \sum_{i=1}^{N_t} \omega_{i,t} * CI_{i,t}^{cs} \tag{4}$$

$$CF_{I,t}^{cs} = \sum_{i=1}^{N_t} \omega_{i,t} * CF_{i,t}^{cs} \tag{5}$$

where N_t is the number of constituents in the respective index at time t. Thus, $\mathrm{CI}_{I,t}^{cs}$ corresponds to the standard weighted-average carbon intensity (WACI) of the index, and $\mathrm{CF}_{I,t}^{cs}$ measures the index's aggregated carbon footprint per unit of market value. In a robustness test, I also use enterprise value including cash (EVIC) instead of market value as the denominator. These carbon metrics closely follow the taxonomy used in TCFD-aligned reporting, and MSCI (2024a) shows that similar WACI and financed-emissions-intensity measures are standard in major providers' carbon-footprinting and climate-risk methodologies.

2.2.2 Data coverage

Refinitiv has gradually expanded its emissions coverage over time. Figure 1 plots the time series of reported carbon-data coverage for both parent indices from December 2005 to December 2023, separately by number of constituents (top panels) and by index weight (bottom panels), and for Scope 1–2 and Scope 1–3. Coverage is especially low at the beginning, with only about 40% of STOXX Europe 600 constituents and around 30–35% of STOXX Europe TMI constituents reporting Scope 1–2 emissions. Since smaller stocks are less likely to disclose their carbon values, coverage by index weight is a lot higher than coverage by number of firms.

Based on this pattern, I restrict the main portfolio analysis to 2011–2024. With the one-year reporting lag described in Section 2.2.1, the first optimization at the end of 2010 uses emissions reported for 2009–2010, when reported Scope 1–2 (Scope 1–3) coverage by index weight already exceeds 80% (60%) in both universes. Reported Scope 1–3 coverage is systematically lower and lags Scope 1–2 by about 10–20 percentage points for much of the sample, reflecting the later and less consistent adoption of Scope 3 reporting. Consequently, I treat Scope 1–2 metrics as the primary target of decarbonization in the main analysis and use Scope 1–3 metrics mainly for robustness.

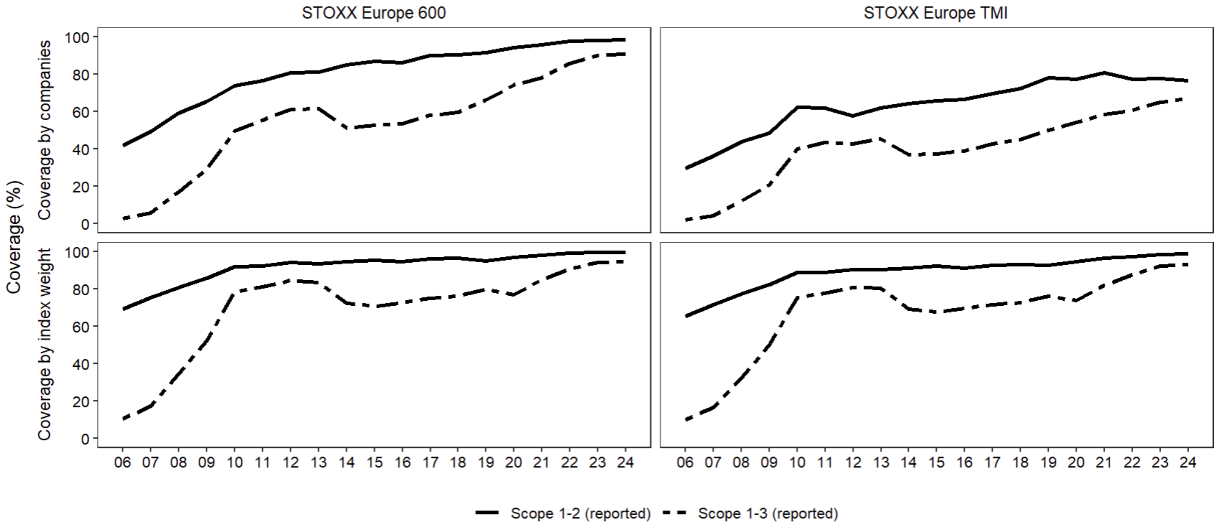


Figure 1. Coverage of reported emissions by scope for STOXX Europe 600 and STOXX Europe TMI, December 2005–December 2023

Notes: This figure plots the time series of reported emissions coverage for Scope 1–2 (solid line) and Scope 1–3 (dashed line) for the STOXX Europe 600 (left panels) and STOXX Europe TMI (right panels). The top panels show coverage by number of index constituents, and the bottom panels show coverage by index weight, measured each December as the percentage of constituents or index weight with non-missing emissions in Refinitiv.

2.2.3 Treatment of missing emissions and benchmark carbon series

Even after 2009, a fraction of index constituents lack reported emissions in Refinitiv, especially for Scope 3. Because I need index-level carbon metrics for the benchmark, I must approximate the carbon contribution of these firms. For the main analysis, I rely on a simple partial-coverage scaling procedure. For each year and ICB sector, I compute sector-level CI and CF using firms with reported Scope 1–2 emissions and rescale these values to the full sector weight in the index, so that non-reporters[5] effectively inherit the sector average. This approach is conceptually similar to the way major index providers obtain near-complete coverage by combining reported emissions with industry-based estimates for non-reporters (MSCI, 2024b). The resulting reported (and sector-scaled) series is my benchmark for Scope 1–2 CI and CF, and I apply the same scaling logic to Scope 1–3.

As a robustness check, I also construct alternative benchmark series based on Refinitiv's model-based Scope 1–2 estimates (LSEG, 2024) and on a Heckman-type selection model that imputes missing emissions.[6] These variants and the corresponding index-level CI and CF paths are reported in Appendix B and have very similar results for Scope 1–2, with correlations to the

[5] By index weight, Scope 1–2 non-reporters account for about 8% of the STOXX Europe 600 and 11% of the TMI at the start of the sample (2009), declining to below 1% by 2023 (see Figure 1).

[6] In line with Olesiewicz et al. (2023), I treat disclosure as missing not at random. For each scope, I estimate a two-equation Heckman selection model: a probit first stage for the reporting decision (as a function of firm size, sector, country, and year) and a log-linear second stage for emissions that includes the inverse Mills ratio from the first stage. I use the fitted second-stage equation to generate single-point estimates for non-reporting firms, rather than the full multiple-imputation procedure proposed by Olesiewicz et al. (2023).

baseline that are close to 1. Therefore, I only use the reported sector-scaled benchmark in the main analysis, as it is transparent and straightforward to implement in practice. Importantly, the input for the optimization is restricted only to reporting firms.

2.3 Firm data

I obtain daily total return data, free-float, and market value for all stocks that ever belong to the STOXX Europe 600 or STOXX Europe TMI from Refinitiv. All returns and prices are expressed in euros. Where necessary, local-currency data are converted using end-of-day exchange rates. Monthly returns are computed from daily returns and are later used for both performance evaluation and the estimation of return covariances.[7] In addition to the carbon variables described in Section 2.2, I collect firm-level accounting data such as revenue, total assets, earnings before interest and taxes (EBIT) and common book equity, which are used to construct factor characteristics and risk-model inputs in Section 3.

Table 1 summarizes firm-level characteristics for the STOXX Europe 600 and STOXX Europe TMI over 2011–2024. Firms in the STOXX Europe 600 are much larger on average, with median free-float market capitalization and revenue roughly three times those in the broader TMI universe. Carbon intensities and carbon footprints are highly skewed in both indices, and somewhat more dispersed in the TMI, reflecting the presence of many smaller and more emission-intensive firms.

3 Method

3.1 Optimization problem and objective

The aim of this paper is to construct low-carbon variants of the STOXX Europe 600 and the STOXX Europe Total Market, that closely track their parent index while reducing the portfolio-level carbon emissions. To achieve this, I solve the following optimization problem for the respective carbon metric CM at each annual rebalancing date t (end of December):

$$\min_{\omega_{P,t}} \sum_{i} \omega_{P,t,i} CM_{i,t}^{cs} \tag{6}$$

where $\omega_{P,t,i}$ is the portfolio weight of stock i at time t of rebalancing. For each stock i, $CM_{i,t}^{cs}$ is the firm-level carbon metric at time t, based upon the chosen set of scopes (see Section 2).

[7] To ensure the quality of the daily return data, I follow Landis and Skouras (2021) filters and drop stocks with implausible return patterns: (i) at least 98% positive or negative returns, (ii) more than 95% zero returns, (iii) unrealistically high (> 40%) or near-zero (< 0.0001%) daily return volatility, and (iv) extreme one-day jumps that reverse the next day (returns ≥ +100% followed by ≤ −50%, or vice versa).

Table 1. Firm-level characteristics by universe, January 2011–December 2024

Panel A: STOXX Europe 600

Variable	Unit/scaling	N	P5	P20	Median	P80	P95
Monthly Return	% per month	100162	-12.8%	-5.2%	0.8%	6.7%	13.8%
Market Value (MV)	€m	100162	1907	3394	7368	21334	63042
Free Float MV	€m	100162	1434	2432	5087	15967	50796
Revenue (Rev)	€m	99234	447	1596	5492	20792	65217
Scope 1	tCO_2e	83963	326	4001	45139	629000	16844500
Scope 2	tCO_2e	82099	902	8000	66200	444000	2581000
Scope 3	tCO_2e	68629	1654	14783	321700	8666214	105000000
Scope 1–2	tCO_2e	89613	1840	15257	138376	1295235	19882066
Scope 1–3	tCO_2e	68221	7782	57006	694048	13190000	125308108
CF Scope 1–2	tCO_2e / MV in €m	81630	0.31	2.41	14.66	139.24	1174.43
CF Scope 1–3	tCO_2e / MV in €m	64696	1.01	6.33	60.23	905.08	6160.76
CI Scope 1–2	tCO_2e / Rev in €m	89200	0.68	4.87	24.36	136.37	939.17
CI Scope 1–3	tCO_2e / Rev in €m	67918	2.19	14.94	152.93	1153.62	4498.13

Panel B: STOXX Europe TMI

Variable	Unit/scaling	N	P5	P20	Median	P80	P95
Monthly Return	% per month	224134	-14.8%	-6.0%	0.6%	7.2%	16.2%
Market Value (MV)	€m	224134	198	705	2337	8841	36663
Free Float MV	€m	224134	91	407	1444	6014	28454
Revenue (Rev)	€m	220657	76	402	1828	8899	39993
Scope 1	tCO_2e	148830	110	1668	21203	345000	9560000
Scope 2	tCO_2e	143789	248	2978	30287	272735	1588000
Scope 3	tCO_2e	117896	655	8000	186891	4203000	56762810
Scope 1–2	tCO_2e	159430	535	6475	67758	709000	10620000
Scope 1–3	tCO_2e	116810	3363	29380	401000	6352083	72791900
CF Scope 1–2	tCO_2e / MV in €m	141540	0.32	2.54	17.38	167.39	1475.23
CF Scope 1–3	tCO_2e / MV in €m	106183	1.04	7.14	77.06	1125.22	7008.49
CI Scope 1–2	tCO_2e / Rev in €m	158535	0.66	4.41	22.80	134.12	982.94
CI Scope 1–3	tCO_2e / Rev in €m	116191	2.56	16.75	171.35	1093.43	4457.69

Notes: This table reports firm-month characteristics for constituents of the STOXX Europe 600 Index (Panel A) and the STOXX Europe Total Market Index (Panel B). For each variable, I report the unit of measurement, the number of firm-month observations (N), and the 5th (P5), 20th (P20), 50th (median), 80th (P80), and 95th (P95) percentiles of the cross-sectional distribution. Emissions are measured in metric tons of CO_2 equivalent (tCO_2e). Carbon footprint (CF) is defined as Scope 1–2 or Scope 1–3 emissions scaled by market value (in €m), and carbon intensity (CI) as emissions scaled by revenue (in €m). Scope 1–2 emissions equal the sum of Scope 1 and Scope 2 where a breakdown is available or the combined Scope 1–2 figure when only an aggregate is reported, so the number of Scope 1–2 observations can exceed that for Scope 1 and Scope 2 individually. The number of observations varies across variables because of missing financial or carbon data and changing index membership over time. Monthly returns are total equity returns in percent per month.

The metric $CM_{i,t}^{cs}$ is either the revenue-based carbon intensity $CI_{i,t}^{cs}$ or the market-value based carbon footprint $CF_{i,t}^{cs}$ as defined in equations (2) and (3).

My minimization is subject to a tracking error TE budget, as well as a set of investability and exposure constraints (see Section 3.2). The predicted ex-ante tracking error of the optimized portfolio relative to its parent index is:

$$TE_t(\omega_{P,t}) = \sqrt{(\omega_{P,t} - \omega_{I,t})^{\mathrm{T}} \Sigma_t (\omega_{P,t} - \omega_{I,t})} \tag{7}$$

where $\omega_{I,t}$ is the vector of index (I) weights at time t and Σ_t is the covariance matrix of stock returns implied by the self-constructed multi-factor risk model at that date (see Online Appendix B for the detailed construction). In the optimization, I impose the following constraint:

$$TE_t\left(\omega_{P,t}\right) \leq TE^* \tag{8}$$

with TE^* being the (annualized) ex-ante tracking-error budget. The remaining constraints are described in the following Section 3.2.

3.2 Investable universe and constraint set

The investable universe at the annual rebalancing at the end of each year consists of all constituents from the respective STOXX index. Starting from the full index membership, I only keep those stocks for which I have non-missing and non-negative carbon data (see Section 2 for data coverage). Within this investable universe, I impose constraints that closely follow those in Bouchey et al. (2024) and low-carbon index methodologies (MSCI, 2024c) to preserve benchmark characteristics and ensure realistic implementability.

First, the portfolio is long-only and fully invested. All portfolio weights $\omega_{P,i,t}$ are restricted to be non-negative and must sum to one:

$$\omega_{P,i,t} \geq 0 \;\forall\, i, \quad \sum_i \omega_{P,i,t} = 1 \tag{9}$$

Second, to avoid extreme weight tilts into small or illiquid stocks, I limit the portfolio weight as a multiple of the benchmark weight:

$$\omega_{P,i,t} \leq L * \omega_{I,i,t} \;\; \forall\, i \tag{10}$$

where $L \geq 1$ is a liquidity multiple. For example, $L = 2$ would allow a stock with a 2% benchmark weight to have at most a 4% weight in the optimized portfolio.

Third, to keep position sizes realistic, so that no single low-carbon stock dominates the entire portfolio, I constrain each stock's weight $\omega_{P,i,t}$ relative to the largest constituent in the benchmark $\omega_{I,t}^{max}$:

$$\omega_{P,i,t} \leq a * \omega_{I,t}^{max} \;\; \forall\, i \tag{11}$$

where $a \geq 1$ is a scalar. For example, at $a = 1.4$ with a max benchmark weight of 5%, this caps any single optimized stock weight at 7%.

To maintain the macro composition of the benchmark, I require country and sector weights to stay within narrow bands around their benchmark levels. For each country c, I impose

$$\left| \sum_{i \in C} \omega_{P,i,t} - \sum_{i \in C} \omega_{I,i,t} \right| \leq \delta_{country} \tag{12}$$

and analogously for each sector s

$$\left| \sum_{i \in S} \omega_{P,i,t} - \sum_{i \in S} \omega_{I,i,t} \right| \leq \delta_{sector} \tag{13}$$

In the most restrictive specification, I set $\delta_{country} = \delta_{sector} = 0.01$, that is, absolute deviations of at most 1 percentage point of index weight.

I also constrain the portfolio's exposures to six standardized style factors (Size, Value, Quality, Investment, Momentum, and Volatility) to ensure that the portfolio's performance is not driven by unintended factor bets. For example, relatively high momentum could unintentionally dominate the portfolio's risk and cause its returns to drift away from the benchmark. For each style factor k, I impose:

$$\left| \sum_{i} \omega_{P,i,t} x_{i,k} - \sum_{i} \omega_{I,i,t} x_{i,k} \right| \leq \delta_{factor} \tag{14}$$

where $\sum_i \omega_{P,i,t} x_{i,k}$ is the portfolio's factor exposure weighted by each stock in the portfolio $\omega_{P,i,t}$ for each style factor $x_{i,k}$, and $\sum_i \omega_{I,i,t} x_{i,k}$ is the benchmark's exposure. All factor exposures are cross-sectionally standardized to mean zero and unit variance each month, so δ_{factor} is expressed in z-score units. For example, $\delta_{factor} = 0.01$ allows only modest deviations from the benchmark along each style factor.

If the optimization problem is infeasible for the baseline parameter values, I iteratively relax the constraints. In each step, I increase the tracking error budget TE^* by 0.0005, raise both the liquidity multiple L and the maximum weight scalar a by 0.1, and increase the country, sector, and style factor exposure bands $\delta_{country}$, δ_{sector}, and δ_{factor} by 0.001, until a feasible solution is obtained.

3.3 Rebalancing and trading mechanism

The backtests cover the period from January 2011 to December 2024 (see Section 2), with annual rebalancing at the end of each calendar year. At each rebalancing date, I construct the investable universe, estimate the risk model (Online Appendix B), and solve the optimization problem (Section 3.1) under constraints (Section 3.2). The optimized portfolio is then held as a buy-and-hold portfolio until the next rebalancing date. The individual optimized weights only evolve through realized total returns, and no trading takes place between scheduled rebalancing dates, so all turnover is concentrated at the annual rebalancing.

Corporate actions such as stock splits or cash dividends are handled by the total return series provided by Refinitiv. New index constituents, including IPOs or stocks created through spin-offs, can only enter the investable universe at the next scheduled rebalancing date. Stocks that leave the parent index or are delisted during the holding period are also removed from the optimized portfolio at their last available price. The returns realized up to this point are effectively held in cash with a return of 0% until the next rebalancing date. With this setup, I mimic a realistic index-tracking implementation with periodic rebalancing, which does not continuously respond with trades to every index inclusion or deletion.

3.4 Scenario design

3.4.1 Main Scenarios

To analyze the trade-off between carbon reduction and active risk, I construct three scenario portfolios that differ in their constraints and tracking-error (TE) budgets, namely, Conservative, Balanced and Aggressive. In their base configuration, all three scenarios use the STOXX Europe 600 as the parent index and apply the optimization framework described in the previous sections. Unless stated otherwise, the objective is to minimize the main measure Scope 1–2 carbon intensity. The corresponding carbon footprint will always be reported additionally.

For the Conservative portfolio, country, sector, and factor exposures are set to 1%, the liquidity multiple is 2x, and the capped weight scalar is 1.2. The Balanced portfolio relaxes the exposures to 2%, the liquidity multiple to 5x and the scalar to 1.5. The Aggressive portfolio allows exposures up to 3%, a liquidity multiple of 10x, and a scalar of 2. Table 2 summarizes the full constraint settings for all three scenarios.

Table 2. Specification of low-carbon portfolio scenarios

Scenario	TE-Budget	Weight restriction		Exposures		
		Liquidity Multiple	Max weight scalar	Country	Sector	Factor
Conservative	0.5%	2	1.2	0.01	0.01	0.01
Balanced	1.0%	5	1.5	0.02	0.02	0.02
Aggressive	2.0%	10	2.0	0.03	0.03	0.03

Notes: This table summarizes the key constraint settings for the three optimized low-carbon portfolios. TE-Budget is the annualized ex ante tracking-error constraint relative to the parent index. The Liquidity Multiple limits each stock's portfolio weight to a multiple of its benchmark weight, and the Max weight scalar caps any stock's weight as a multiple of the largest benchmark constituent. The Country, Sector, and Factor columns report the absolute bands on deviations of portfolio country and sector weights (in percentage points of index weight) and standardized style-factor exposures from their benchmark levels.

The TE budget for each portfolio is chosen using the carbon frontier shown in Figure 2.[8]

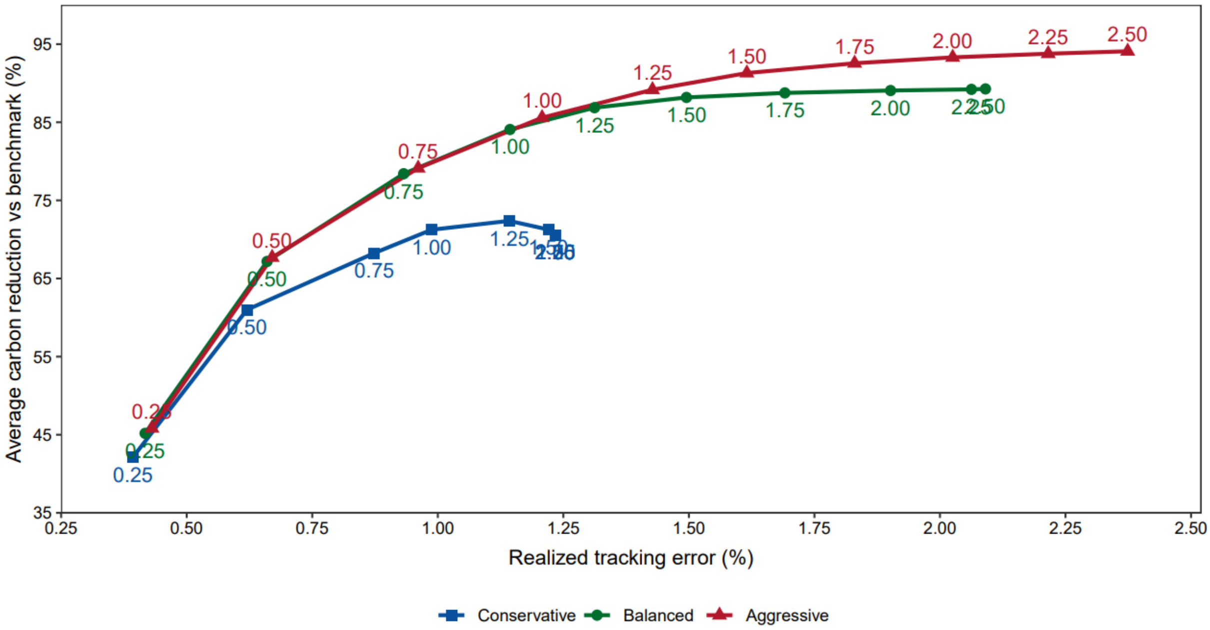


Figure 2. Carbon–tracking error frontier for Scope 1–2 intensity: STOXX Europe 600, January 2011–December 2024

Notes: This figure plots the average percentage reduction in portfolio Scope 1–2 carbon intensity relative to the STOXX Europe 600 benchmark against realized tracking error for three constraint sets (Conservative, Balanced, Aggressive). For each scenario, portfolios are optimized to minimize revenue-based Scope 1–2 carbon intensity under a grid of ex ante tracking-error budgets from 0.25% to 2.50%, holding all other constraints fixed as specified in Table 2. Points along each line correspond to different ex ante tracking-error budgets, which are shown as labels next to the markers. Realized tracking error and average carbon reductions are computed from monthly returns over 2011–2024.

The figure plots the average carbon reduction relative to the benchmark against realized TE for a grid of ex-ante TE budgets between 25 and 250 basis points. The carbon frontier shows strongly diminishing marginal reduction beyond 1% to 2% TE, depending on the scenario (Conservative, Balanced, and Aggressive). Based on this frontier, I select representative TE

[8] The frontier is inspired by Exhibit 8 (p. 68) from Bouchey et al. (2024).

budgets of 0.5% for the Conservative portfolio, 1.0% for the Balanced portfolio, and 2.0% for the Aggressive portfolio. These points correspond to portfolios that are, respectively, very close to the benchmark, moderately active, and strongly tilted toward low-carbon stocks, while covering the range of TE levels commonly observed in practice.[9] Section 4.1 reports performance, risk, and carbon outcomes for these three main scenarios.

3.4.2 Robustness specifications

In additional specifications, I vary the carbon objective, the emissions scope, the parent index, the sample period, the valuation base in the carbon footprint denominator, and the investor's reporting currency. The goal is to assess whether the main findings for my three scenario portfolios optimized for low Scope 1–2 carbon intensity are robust across these design choices.

First, I replace the low Scope 1–2 carbon intensity objective with Scope 1–2 carbon footprint, as well as a variant which uses EVIC instead of market value as the denominator.[10]

Second, I then repeat this analysis with the broader Scope 1–3 carbon intensity and separately with Scope 1–3 carbon footprint. For each of these objectives, I keep the constraint sets and TE budgets identical to the main scenarios and compare realized TE, the achieved reduction in the carbon intensity and footprint, and basic performance statistics. These variations summarize how sensitive my results are to using absolute versus intensity-based objectives and to including value-chain emissions.

Third, I modify the investment universe and time period. I rerun the optimization with the STOXX Europe Total Market as the parent index to examine whether the results depend on the breadth of the universe. In addition, I winsorize firm-level Scope 1–2 carbon intensities at 1-5% two-sided percentile cut-offs to verify that the findings are not driven by a few extreme observations. I also estimate portfolios which start post-Paris in January 2016, when climate policies and investor awareness increased, and in January 2020, to see if similar carbon reductions are still possible in more recent years.

Fourth, I analyze broader implementation, sustainability, and factor exposure effects. For each scenario portfolio, I calculate annual turnover and net performance under realistic

[9] For example, the MSCI World Screened Index has a TE of 0.66%, while the MSCI World Climate Change Index has a TE of 2.00% according to their latest index factsheets (as of Oct 2025).

[10] EVIC is computed as enterprise value (EV, WC18100) plus a sector-specific (ICBSN) cash component: Cash & Equivalents Generic (WC02005) for life and non-life insurers, Cash & Due from Banks (WC02004) for banks, finance and credit services, and investment banking and brokerage, and Cash (WC02003) for all other firms (all using lagged values). The usage of EVIC has become widespread because it can be applied to multi-asset portfolios (MSCI, 2024a).

transaction cost and fee assumptions to approximate implementable returns.[11] I then assess whether the optimization unintentionally worsens the environmental profile by comparing Refinitiv's ESG, environmental, resource use, environmental innovation, and ESG combined scores, together with MSCI's ESG and environmental scores.[12] Lastly, I analyze the ex-post exposures of the portfolios to the six Fama–French factors to verify that the carbon optimization does not inadvertently introduce systematic factor tilts that could confound the performance results.[13]

Finally, I adopt the perspective of a US investor in the European equity market and recompute all portfolio returns and performance metrics in US dollars.[14]

3.5 Evaluation Metrics

I evaluate each portfolio along three dimensions that reflect the main investor objectives in low-carbon portfolio design: carbon reduction, tracking quality, and risk-return trade-offs.

Carbon performance is summarized by the reduction in Scope 1–2 and Scope 1–3 carbon intensity and carbon footprint relative to the parent index. For each year t, I compute the difference between the optimized portfolio's index-level CI_t^{cs} and CF_t^{cs} and those of the benchmark (equations 4–5), express this difference as a percentage of the benchmark level, and report time-series averages of these percentage reductions over the sample period.

Tracking quality is measured by the realized (ex post) tracking error, defined as the standard deviation of monthly active returns of the optimized portfolio relative to its parent index, annualized by multiplying by $\sqrt{12}$. This ex post TE is different from the ex ante TE constraint used in the optimization problem (equation 8).

Risk-return performance is summarized by annualized mean return, annualized volatility, and the Sharpe and Information ratios. In Section 4, I report these metrics for the benchmark and for the Conservative, Balanced, and Aggressive portfolios. In addition, I decompose cumulative active returns into sector allocation, within-sector selection, and interaction

[11] I assume an annual management fee of 20 bps and proportional transaction costs of 25 bps per 100% of portfolio turnover. These costs are broadly in line with fee and trading-cost levels for large low-carbon or ESG equity ETFs.
[12] I use environmental and ESG scores from two providers to account for known provider differences (Berg et al., 2022).
[13] The six Fama–French factors are constructed from STOXX Europe TMI constituents using the standard 2×3 sorts on size and book-to-market, profitability, investment, and prior 2–12-month momentum. Portfolios are value-weighted, rebalanced monthly, and factor returns are formed as in Fama and French (2015, 2018). I construct these factors myself rather than using the Kenneth R. French Data Library, because the latter provides European factor returns in USD, whereas my portfolio and benchmark returns are EUR-denominated.
[14] All inputs, including individual stock returns, factor returns, and the variance-covariance matrix as well as all risk model parameters, are also calculated using US dollar returns.

components using a standard Brinson attribution across ICB sectors (Brinson, Hood, and Beebower, 1986), which I report in Table 5.

4 Results

4.1 Performance and risk of three main scenarios

Table 3 reports carbon, risk, and performance outcomes for the Conservative, Balanced, and Aggressive portfolios that minimize Scope 1–2 carbon intensity relative to the STOXX Europe 600 over 2011–2024.

Table 3. Performance and carbon outcomes for main low-carbon portfolio scenarios: STOXX Europe 600, January 2011–December 2024

Metric	Benchmark	Conservative	Balanced	Aggressive
ex-ante Tracking Error budget (%)	-	0.50	1.00	2.00
ex-post Tracking Error realized (%)	-	0.64	1.17	2.08
Carbon Intensity Reduction (S1–2, %)	**0.0**	**61.0**	**84.1**	**93.3**
Carbon Footprint Reduction (S1–2, %)	0.0	57.5	79.3	93.3
Carbon Intensity Reduction (S1–3, %)	0.0	26.1	48.2	66.3
Carbon Footprint Reduction (S1–3, %)	0.0	20.8	40.3	64.5
Cumulative return (%)	196.1	209.2	230.8	267.5
Annual return (%)	8.06	8.40	8.92	9.74
Annual volatility (%)	13.31	13.31	13.26	13.37
Sharpe Ratio	0.61	0.63	0.67	0.73
Information Ratio	-	0.52	0.73	0.81

Notes: This table reports tracking-error constraints, carbon reductions, and performance statistics for three optimized low-carbon portfolios (Conservative, Balanced, Aggressive) and the STOXX Europe 600 parent index. Portfolios minimize revenue-based Scope 1–2 carbon intensity subject to ex ante tracking-error budgets and the investability and exposure constraints described in Section 3, and are rebalanced annually each December. Carbon intensity and carbon footprint reductions for Scope 1–2 and Scope 1–3 are expressed as percentage reductions relative to the benchmark. Cumulative and annual returns, volatility, and Sharpe and information ratios are based on monthly total returns in euros, gross of management fees and transaction costs; Appendix Table C.6 reports net-of-cost results.

The key result for investors interested in low-carbon portfolios is that substantial Scope 1–2 reductions are possible even at low to moderate levels of tracking error. For an ex ante TE budget of 0.5% (Conservative), the optimized portfolio achieves an average Scope 1–2 carbon intensity reduction of about 61% and a 58% reduction in Scope 1–2 carbon footprint, at a realized TE of 0.64%. The Balanced scenario (1% TE budget) lowers Scope 1–2 intensity by 84% and footprint by 79%, for a realized TE of 1.17%. In the Aggressive scenario (2% TE budget), Scope 1–2 intensity and footprint reductions exceed 93%, while ex post TE remains close to the budget (2.08%). The marginal gain in carbon reduction per unit TE, however, strongly levels off. Moving from Conservative to Balanced adds more than 20 percentage points of intensity reduction for roughly half a percentage point of extra TE, whereas going from

Balanced to Aggressive yields only about 9 percentage points of additional reduction for almost 1 percentage point of extra TE.

Although the portfolios are optimized only on Scope 1–2 intensity, they also substantially reduce Scope 1–3 metrics. Scope 1–3 carbon-intensity reductions range from about 26% (Conservative) to 66% (Aggressive), and Scope 1–3 carbon-footprint reductions from roughly 21% to 65%. Thus, targeting direct and indirect emissions (Scope 1–2) already yields sizeable co-benefits for value-chain emissions reductions (Scope 1–3).

Tracking quality remains high across all three scenarios. Ex post TE is close to the ex ante budgets and relatively low in absolute terms, between 0.64% and 2.08% per year. Total volatility of the optimized portfolios is virtually identical to that of the benchmark (around 13.3% per year), indicating that decarbonization does not lead to materially riskier portfolios in terms of total risk.

Importantly, there is no evidence of a performance penalty associated with decarbonization in this sample. Annualized returns increase from about 8.1% for the benchmark to 8.4%, 8.9%, and 9.7% for the Conservative, Balanced, and Aggressive portfolios, respectively, leading to cumulative return differences of roughly 13–71 percentage points. Given that volatility remains similar, the Sharpe ratio rises slightly from 0.61 to 0.63, 0.67, and 0.73 across the three scenarios. In benchmark-relative terms, the information ratio is positive for all three portfolios and increases with the TE budget, from 0.52 in the Conservative case to 0.73 and 0.81 in the Balanced and Aggressive cases. For net values, see Appendix Table C.6.

Overall, the Balanced portfolio appears particularly attractive for investors, as it captures most of the achievable reduction while keeping tracking error, constraint relaxations, and deviations from the benchmark moderate.

4.2 Robustness of main scenarios

Appendix Tables C.1 to C.7 test the robustness of the main findings with respect to the carbon objective and denominator, the emissions scope, the investment universe, the sample period, and the investor's reporting currency. Across these variations, the main message is that large emissions reductions are generally achievable at low to moderate tracking error and that risk-return profiles remain comparable to, or slightly better than, those of the parent index.

Alternative objectives and denominators (Appendix Table C.1).

When Scope 1–2 carbon footprint replaces Scope 1–2 intensity as the objective, the optimized portfolios deliver similar reductions in the targeted footprint metric (Panel A).

Depending on the tracking-error budget, Scope 1–2 footprint falls by 61 to 97 percent, while Scope 1–2 intensity still declines by around 50 to more than 88 percent. Realized tracking errors are somewhat higher than in the main scenarios, especially for the more aggressive portfolios, and the two lower TE portfolios show a small performance drag relative to the benchmark. Using EVIC instead of market value in the footprint denominator leaves the qualitative picture unchanged (Panel B). Thus, the main decarbonization results are not driven by the choice between intensity and footprint, or by the specific valuation base in the denominator, although a footprint-only objective at tighter TE budgets can slightly worsen benchmark-relative performance.

Broader emissions scopes (Appendix Table C.2).

Extending the objective to Scope 1–3 metrics broadly confirms the main results but slightly weakens the Scope 1–2 gains and raises realized TE somewhat. Minimizing Scope 1–3 carbon intensity cuts Scope 1–3 intensity by about 29, 67 and 94 percent, while Scope 1–2 intensity still falls by roughly 28, 63 and 85 percent (Panel A). Although those reductions are still substantial, they are smaller than in the baseline Scope 1–2 optimization. Information ratios, however, remain positive. When Scope 1–3 footprint is minimized instead, footprint reductions reach around 26, 73 and 98 percent, again with sizeable but smaller Scope 1–2 improvements, but the low-TE portfolios now slightly underperform the benchmark, and only the high-TE portfolio combines very large Scope 1–3 footprint cuts with a positive information ratio (Panel B).

Alternative parent index, winsorization and time windows (Appendix Tables C.3 to C.5).

Using the broader STOXX Europe TMI as the parent index (Appendix Table C.3) results in very similar patterns to the STOXX Europe 600. Scope 1–2 carbon intensity and footprint can again be reduced by roughly 64–94 percent across the three scenarios, while realized tracking errors remain close to their ex ante budgets, and information ratios are positive and increasing with the TE budget. In a further check, I winsorize Scope 1–2 carbon intensities at different two-sided percentile cutoffs before running the optimization (Appendix Table C.4). Moving from the baseline (no winsorization) to winsorization at the 1–5 percent levels gradually lowers the reported decarbonization figures, but the Conservative, Balanced, and Aggressive portfolios still achieve around 51, 77, and 90 percent lower Scope 1–2 intensity than the benchmark at the 5 percent winsorization level. This indicates that the main findings are not driven by a few extreme carbon observations. Restricting the sample to the post-Paris period (2016–2024) or to the very recent years 2020–2024 (Appendix Table C.5) likewise leaves the main conclusions unchanged. Scope 1–2 reductions remain in a similar range and ex post TE stays close to the budget. All three portfolios outperform the benchmark on a risk-adjusted basis. However, the Conservative portfolio now shows the highest and Aggressive the lowest information ratio.

Implementation, sustainability, and factor exposures (Appendix Table C.6).

Appendix Table C.6 Panel A (implementation and trading characteristics) shows that the three optimized portfolios remain implementable and, even after costs, do not give up performance. Cumulative net returns range from 196.9% (Conservative) to 252.4% (Aggressive), so the portfolios still outperform the benchmark after fees and transaction costs. Net information ratios remain positive (0.03, 0.45, 0.65). Average active weight increases from about 22% to 62% from Conservative to Aggressive, while annual turnover stays in a moderate 45–53% range. The number of holdings falls with more aggressive tilts, but even the Aggressive portfolio still holds around 160 stocks, indicating that portfolios remain well diversified. Panel B (sustainability performance) suggests that decarbonization does not unintentionally weaken the broader ESG profile. For Refinitiv, ESG combined[15], ESG, resource use, and the environmental pillar are generally slightly higher than for the benchmark; only the Aggressive portfolio shows a small dip in two Refinitiv sustainability scores relative to the index. For MSCI, both the ESG and environmental pillar scores are modestly higher for all three optimized portfolios. Overall, the sustainability profile is at least preserved and often marginally improved. Panel C (factor exposures) reports differences in Fama–French six-factor betas relative to the benchmark. The Conservative portfolio is virtually indistinguishable from the index. The Balanced and Aggressive portfolios exhibit very moderately more negative HML (value) and the latter exhibits higher momentum loadings, with other factor deviations small in magnitude.

US investor perspective (Appendix Table C.7).

Finally, I recompute all portfolio and benchmark returns in US dollars to mimic the perspective of a US investor in European equities. The resulting carbon reductions, realized tracking errors, and relative performance statistics[16] are very similar to those in euro terms. This confirms that the main conclusions are not sensitive to the reporting currency.

A common finding across nearly all specifications is that the optimized low-carbon portfolios generate positive information ratios. Several forces contributed to this pattern. First, low-carbon stocks themselves outperformed their high-carbon counterparts. Simple value-weighted portfolios sorted on monthly lagged carbon intensity show that low-CI stocks (bottom

[15] The Refinitiv datapoint "ESG Combined" consists of the weight-aggregated three pillar score ESG, and the Controversy Score.
[16] Absolute performance is worse due to the Euro's depreciation versus the US dollar over the period.

30% or 10%) outperformed high-CI stocks (top 30% or 10%) by approximately 40–60 percentage points cumulatively, with the spread widening for more extreme sorts (Appendix D). Second, the Balanced and especially the Aggressive portfolios tilted modestly towards momentum and away from value (Appendix Table C.6 Panel C). Both tilts were rewarded during the sample period, MSCI Europe Momentum outperformed while MSCI Europe Value lagged the broader market. Since these tailwinds operated in the same direction, it is difficult to isolate a pure "carbon alpha" from slightly favorable factor dynamics. Whether the observed outperformance would thus persist in a value-led or momentum-reversing environment remains an open question for future research.

4.3 Sector allocations and within-sector stock selection

Table 4 summarizes how the optimized portfolios deviate from the STOXX Europe 600 at the sector level. Panel A reports the average active weights. The pattern largely reflects how the optimizer uses the sector bands: sectors with high benchmark carbon intensity such as Utilities and Basic Materials are typically pushed toward the lower end of their allowed range (-1 to -3%), while low-intensity sectors such as Financials and Telecommunications are moved toward the upper end of their ranges (+1 to +3%).[17] Panel B shows that these relatively small sector tilts are accompanied by substantial within-sector reshuffling. For the Conservative portfolio, roughly 10%–30% of each sector's weight is reallocated away from the benchmark holdings, while in the Balanced and Aggressive portfolios the within-sector shifts reach about 30%–60% of sector weights in most industries.

Table 4. Sector tilts and within-sector rebalancing for low-carbon portfolios: STOXX Europe 600, January 2011–December 2024

Panel A: Sector active weights relative to the benchmark (percentage points)

Sector	Conservative	Balanced	Aggressive
Financials	0.83	1.88	3.00
Telecommunications	0.76	1.37	1.92
Health Care	0.62	1.29	1.68
Industrials	0.19	0.33	0.08
Consumer Staples	0.06	-0.25	-2.74
Consumer Discretionary	0.02	0.72	2.82
Technology	0.07	0.30	1.59
Real Estate	-0.14	-0.22	0.42
Energy	-0.44	-1.74	-2.87
Basic Materials	-0.95	-1.67	-2.89
Utilities	-1.00	-2.00	-3.00

[17] Over 2011–2024, median Scope 1–2 carbon intensity is highest for Utilities (141), Basic Materials (68), and Consumer Staples (40), and lowest for Financials (1.8), Technology (7.3), and Consumer Discretionary (8.7).

Table 4. Sector tilts and within-sector rebalancing for low-carbon portfolios: STOXX Europe 600, January 2011–December 2024 (continued)

Panel B: Within-sector active reallocation relative to the benchmark (% of sector weight)

Sector	Conservative	Balanced	Aggressive
Financials	12.8	19.6	36.7
Telecommunications	9.6	19.1	59.6
Health Care	11.1	20.3	39.1
Industrials	26.3	41.8	67.2
Consumer Staples	16.9	37.4	57.4
Consumer Discretionary	19.7	28.7	57.0
Technology	14.9	21.2	46.4
Real Estate	27.3	40.5	58.5
Energy	17.7	31.1	50.2
Basic Materials	30.5	56.2	51.6
Utilities	26.5	36.2	23.0

Notes: This table summarizes how the three optimized low-carbon portfolios (Conservative, Balanced, Aggressive) deviate from the STOXX Europe 600 at the sector level. Panel A reports average sector active weights in percentage points of index weight; positive (negative) values indicate overweights (underweights) relative to the benchmark. Panel B reports within-sector active reallocation, measured as the percentage of each sector's weight that is reassigned across stocks relative to the benchmark, so higher values indicate stronger stock-level reshuffling inside the sector. All figures are averaged over the sample period.

To assess how these sector reallocations translate into carbon outcomes, Table 5 Panel A decomposes the Scope 1–2 carbon-intensity reductions into sector and within-sector components. Panel A shows that the sector tilts account on average for about 15%, 23%, and 33% of the absolute carbon-intensity reduction in the Conservative, Balanced, and Aggressive portfolios, with relatively little time variation. The remaining 67% to 85% of the reduction is attributable to within-sector stock selection. Thus, even for the Aggressive portfolio, the majority of emissions reduction comes from replacing high-emitting stocks with lower-emitting peers inside sectors rather than shifting weights across sectors.

Panel B of Table 5 reports the corresponding decomposition of cumulative active returns over 2011–2024. In the Conservative and Balanced portfolios, within-sector selection contributes 110% and 92% of total active return, with sector allocations contributing slightly negatively (positively) for the former (latter) portfolio. In the Aggressive portfolio, 74% is due to within-sector selection and 21% to sector allocation, with the interaction term small in all three cases.[18]

[18] This pattern contrasts with Bouchey et al. (2024), who report that sector allocation accounts for the majority of both carbon reduction and active return under looser ±5% sector constraints, while within-sector selection contributes negatively to performance (Exhibit 5 and 6, p. 65–66).

Table 5. Sector-based decomposition of carbon-intensity reductions and active returns: STOXX Europe 600, January 2011–December 2024

Panel A: Contribution of sector tilts and within-sector selection to Scope 1–2 carbon-intensity reduction (% of absolute reduction)

Attribution	Statistic	Conservative	Balanced	Aggressive
Sector tilts	Mean	15.4	22.6	33.0
	Min	11.0	16.8	25.8
	Max	21.2	25.5	37.1
Within-sector selection	Mean	84.6	77.4	67.0
	Min	78.8	74.5	62.9
	Max	89.0	83.2	74.2

Panel B: Brinson sector attribution of cumulative active return (percentage points and percent of total active return)

Attribution	Conservative		Balanced		Aggressive	
	ppts	% of active	ppts	% of active	ppts	% of active
Allocation	-0.38	-7.9	1.25	10.4	5.04	21.1
Selection	5.25	109.9	11.02	91.9	17.67	74.0
Interaction	-0.10	-2.1	-0.29	-2.4	1.17	4.9

Notes: This table decomposes carbon-intensity reductions and benchmark-relative performance of the three optimized low-carbon portfolios (Conservative, Balanced, Aggressive) relative to the STOXX Europe 600. Panel A reports the contribution of sector tilts and within-sector stock selection to the absolute reduction in Scope 1–2 carbon intensity, shown as the mean, minimum, and maximum share (in percent of total absolute reduction) across portfolio years. Panel B reports a standard Brinson sector attribution of cumulative active return, splitting the benchmark-relative performance over 2011–2024 into allocation, selection, and interaction effects. For each portfolio, attribution effects are shown in percentage points of cumulative active return (“ppts”) and as a percentage of total cumulative active return (“% of active”). Negative values indicate a drag on active performance.

Combined, Tables 4 and 5 indicate that the optimizer primarily relies on within-sector stock selection to deliver both the strong reductions in portfolio carbon intensity and the modest benchmark outperformance, while sector tilts remain moderate and often operate well inside, rather than at, their allowed bounds.

5 Conclusion

European equity portfolios can reduce Scope 1–2 carbon intensity by 60–90 percent with tracking errors of only 0.5–2 percent while preserving, or modestly improving, risk-adjusted performance. This paper demonstrates this finding by constructing low-carbon variants of the STOXX Europe 600 that minimize portfolio-level carbon metrics subject to tracking-error, sector, country, liquidity, and factor-exposure constraints, using only reported emissions from Refinitiv and a transparent, replicable optimization framework.

The Conservative, Balanced, and Aggressive scenarios reduce Scope 1–2 carbon intensity by roughly 61, 84, and 93 percent for realized tracking errors of 0.64, 1.17, and 2.08 percent per year. Scope 1–2 carbon footprint declines by similar magnitudes, and Scope 1–3 metrics also fall substantially even though they are not targeted directly. Total volatility remains close to the benchmark, and information ratios are positive across all scenarios, ranging from approximately 0.5 for the Conservative portfolio to 0.8 for the Aggressive portfolio. These results are robust to alternative carbon objectives (intensity versus footprint), emissions scopes, denominators (market value vs EVIC), parent indices (including the broader STOXX Europe Total Market), time windows, and reporting currencies. Broader untargeted sustainability metrics, including Refinitiv and MSCI ESG scores, are maintained or modestly improved.

The mechanism underlying these results has direct implications for portfolio construction. Sector deviations remain small, with modest underweights in high-emitting industries and modest overweights in lower-emitting ones. A decomposition of carbon-intensity reductions shows that 67 to 85 percent of the reduction comes from within-sector stock selection rather than sector tilts. A Brinson-style performance attribution yields a similar conclusion, which is that within-sector selection explains the majority of active return, while sector-allocation effects are smaller. This implies that asset managers can use portfolio optimization to maintain sector exposures close to standard benchmarks while tilting toward cleaner stocks within sectors. This approach may be more acceptable to clients and easier to integrate into existing investment processes. Among the three scenarios, the Balanced portfolio appears particularly attractive because it captures most of the attainable emissions reduction with only moderate tracking error and benchmark deviation.

A further practical advantage is that the implementation does not rely on proprietary risk models or vendor-specific optimizers. The approach can be replicated with standard return data and a transparent multifactor covariance model, lowering the barrier for asset managers and asset owners who cannot or do not wish to rely on commercial products.

The results are partially limited by lower reported emissions coverage for smaller firms and for Scope 3. Further, different data providers may also yield different results, although Scope 1–2 data are relatively stable across sources (Heldmann et al., 2025). Finally, the analysis uses annual rebalancing with a simplified transaction-cost specification. A more frequent rebalancing or different cost assumptions could slightly alter net performance results.

For investors seeking to reduce portfolio emissions without departing significantly from benchmark characteristics, a within-sector approach to low-carbon portfolio construction offers a transparent and implementable solution. How such strategies perform in a different carbon-return environment, and how they interact with forward-looking climate scenarios and broader decarbonization dynamics, remains an open question for future research.

References

Alda, M., F. Muñoz, and M. Vargas. 2022. "Product differentiation in the socially responsible mutual fund industry." *Journal of Multinational Financial Management* 64: 100730. https://doi.org/10.1016/j.mulfin.2022.100730.

Ameli, N., P. Drummond, A. Bisaro, M. Grubb, and H. Chenet. 2020. "Climate finance and disclosure for institutional investors: Why transparency is not enough." *Climatic Change* 160 (4): 565–589. https://doi.org/10.1007/s10584-019-02542-2.

Andersson, M., P. Bolton, and F. Samama. 2016. "Hedging climate risk." *Financial Analysts Journal* 72 (3): 13–32. https://doi.org/10.2469/faj.v72.n3.4.

Balp, G., and G. Strampelli. 2022. "Institutional investor ESG engagement: The European experience." *European Business Organization Law Review* 23 (4): 869–904. https://doi.org/10.1007/s40804-022-00266-y.

Bell, M., and B. Taylor. 2024. "How can investors balance short-term demands with long-term value?" EY Insights, December 10, 2024. Accessed July 20, 2025. https://web.archive.org/web/20250725165911/https://www.ey.com/en_gl/insights/climate-change-sustainability-services/institutional-investor-survey.

Bender, J., C. He, C. Ooi, and X. Sun. 2020. "Reducing the carbon intensity of low volatility portfolios." *The Journal of Portfolio Management* 46 (3): 108–122. https://doi.org/10.3905/jpm.2020.46.3.108.

Bender, J., T. A. Bridges, and K. Shah. 2019. "Reinventing climate investing: Building equity portfolios for climate risk mitigation and adaptation." *Journal of Sustainable Finance & Investment* 9 (3): 191–213. https://doi.org/10.1080/20430795.2019.1579512.

Berg, F., J. F. Kölbel, and R. Rigobon. 2022. "Aggregate confusion: The divergence of ESG ratings." *Review of Finance* 26 (6): 1315–1344. https://doi.org/10.1093/rof/rfac033.

Blitz, D., M. Chen, C. Howard, and H. Lohre. 2024. "3D investing: Jointly optimizing return, risk, and sustainability." *Financial Analysts Journal* 80 (3): 59–75. https://doi.org/10.1080/0015198X.2024.2335142.

Bolton, P., M. Kacperczyk, and F. Samama. 2022. "Net-zero carbon portfolio alignment." *Financial Analysts Journal* 78 (2): 19–33. https://doi.org/10.1080/0015198X.2022.2033105.

Bouchey, P., M. de Leon, Z. Jawaid, and V. Nemtchinov. 2024. "Reducing the carbon footprint of an index: How low can you go?" *The Journal of Wealth Management* 27 (2): 57–74. https://doi.org/10.3905/jwm.2024.1.245.

Brinson, G. P., L. R. Hood, and G. L. Beebower. 1986. "Determinants of portfolio performance." *Financial Analysts Journal* 42 (4): 39–44. https://doi.org/10.2469/faj.v42.n4.39.

Busch, T., M. Johnson, and T. Pioch. 2022. "Corporate carbon performance data: Quo vadis?" *Journal of Industrial Ecology* 26 (1): 350–363. https://doi.org/10.1111/jiec.13008.

Cao, S. S., Y. Chen, H. Xiao, and A. L. Zhang. 2025. "The economics of greenwashing funds." SSRN Working Paper. https://ssrn.com/abstract=5040390.

Cenedese, G., S. Han, and M. T. Kacperczyk. 2023. "Carbon-transition risk and net-zero portfolios." SSRN Working Paper. https://doi.org/10.2139/ssrn.4565220.

Cheng, G., E. Jondeau, and B. Mojon. 2022. "Building portfolios of sovereign securities with decreasing carbon footprints." *BIS Working Papers* 1038. Bank for International Settlements, Basel, Switzerland. https://doi.org/10.2139/ssrn.4207316.

Fahlenbrach, R., and E. Jondeau. 2023. "Greening the Swiss National Bank's portfolio." *The Review of Corporate Finance Studies* 12 (4): 792–833. https://doi.org/10.1093/rcfs/cfad011.

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

FTSE Russell. 2022. *Sustainable investment: 2022 global survey findings from asset owners*. FTSE Russell. Accessed April 4, 2026. https://www.lseg.com/content/dam/ftse-russell/en_us/documents/reports/global-asset-owner/sustainable-investment-2022-global-survey-findings-from-asset-owners.pdf.

Giglio, S., M. Maggiori, J. Stroebel, Z. Tan, S. P. Utkus, and X. Xu. 2025. "Four facts about ESG beliefs and investor portfolios." *Journal of Financial Economics* 164: 103984. https://doi.org/10.1016/j.jfineco.2024.103984.

Grinold, R. C., and R. N. Kahn. 2000. *Active portfolio management: A quantitative approach for providing superior returns and controlling risk*. 2nd ed. New York, NY: McGraw-Hill.

Gutsche, G., H. Wetzel, and A. Ziegler. 2023. "Determinants of individual sustainable investment behavior – A framed field experiment." *Journal of Economic Behavior & Organization* 209: 491–508. https://doi.org/10.1016/j.jebo.2023.03.016.

Heeb, F., J. F. Kölbel, F. Paetzold, and S. Zeisberger. 2023. "Do investors care about impact?" *The Review of Financial Studies* 36 (5): 1737–1787. https://doi.org/10.1093/rfs/hhac066.

Heldmann, J., H. D. Dang, and T. Brückner. 2025. "Investing in a low-carbon transition: Carbon footprint savings with green MSCI indices." *Journal of Environmental Management* 393: 126022. https://doi.org/10.1016/j.jenvman.2025.126022.

In, S. Y., K. Y. Park, and A. Monk. 2019. "Is 'being green' rewarded in the market? An empirical investigation of decarbonization and stock returns." *Stanford Global Projects Center Working Paper*. https://www.sustainablefinance.ch/upload/cms/user/SSRN-id3020304.pdf.

Jia, X., Y. Lan, and N. Duong. 2025. "How to decarbonize multi-strategy equity portfolios: Practical considerations." State Street Global Advisors. https://www.ssga.com/de/en_gb/institutional/insights/how-to-decarbonize-multi-strategy-equity-portfolios.

Jondeau, E., B. Mojon, and L. A. Pereira da Silva. 2025. "Building benchmark portfolios with decreasing carbon footprints." *Journal of Sustainable Finance & Investment*. Advance online publication. https://doi.org/10.1080/20430795.2025.2463434.

Kolasa, T., and Z. Sautner. 2025. "Institutional investors and the fight against climate change." *Corporate Governance: An International Review* 33 (4): 663–679. https://doi.org/10.1111/corg.12620.

Kolle, J., H. Lohre, E. Radatz, and C. Rother. 2022. "Factor investing in Paris: Managing climate change risk in portfolio construction." *Journal of Investment Management* 20 (4): 35–51. https://joim.com/factor-investing-in-paris-managing-climate-change-risk-in-portfolio-construction/.

Krueger, P., Z. Sautner, and L. T. Starks. 2020. "The importance of climate risks for institutional investors." *The Review of Financial Studies* 33 (3): 1067–1111. https://doi.org/10.1093/rfs/hhz137.

Landis, C., and S. 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.

Lee, L.-E., T. Séné, R. Mahmood, and U. Ashfaq. 2025. *Transition finance tracker: Third quarter 2025. A quarterly report on financing the shift to a low-carbon economy*. MSCI Institute. Accessed January 29, 2026. https://www.msci-institute.com/wp-content/uploads/2025/10/MSCI-Institute-Transition-Finance-Tracker-Q3-2025.pdf.

LSEG (London Stock Exchange Group). 2024. *LSEG ESG carbon data and estimate models*. London Stock Exchange Group. https://www.lseg.com.cn/content/dam/data-analytics/en_us/documents/fact-sheets/lseg-esg-carbon-data-and-estimate-models.pdf.

Martini, A. 2021. "Socially responsible investing: From the ethical origins to the sustainable development framework of the European Union." *Environment, Development and Sustainability* 23 (11): 16874–16890. https://doi.org/10.1007/s10668-021-01375-3.

Mazzolini, G., D. Salakhova, M. Giuzio, and S. Kapadia. 2025. "Sustainability labels vs. reality: how climate-friendly are green and ESG funds?" ECB Working Paper Series 3121. European Central Bank, Frankfurt am Main, Germany. https://doi.org/10.2139/ssrn.5531680.

Morgan Stanley Institute for Sustainable Investing. 2025. *Sustainable signals: Individual investors 2025*. Morgan Stanley Institute for Sustainable Investing. https://www.morganstanley.com/content/dam/msdotcom/en/assets/pdfs/2025_Sustainable_Signals_Individual_Investors_2025_report.pdf.

MSCI. 2024a. *Carbon Footprinting Demystified: Navigating carbon metrics in a rapidly changing landscape*. MSCI Research Insights. https://www.msci.com/documents/10199/a57d8424-34a9-3ea4-5a74-030349845bb3.

MSCI. 2024b. *Greenhouse Gas Emissions Methodologies*. MSCI ESG Research. https://www.msci.com/documents/1296102/39141520/MSCI%2BGreenhouse%2BGas%2BEmissions%2BMethodologies.pdf/973fb5dc-7ee1-75ac-4c62-a7d36e3b429f.

MSCI. 2024c. *MSCI Global Low Carbon Target Indexes methodology.* MSCI Index Methodology. https://www.msci.com/indexes/documents/methodology/2_MSCI_Global_Low_Carbon_Target_Indexes_Methodology_20241113.pdf.

Olesiewicz, M. P., J. Kooroshy, and S. Greven. 2023. "Navigating the corporate disclosure gap: Modeling of missing not at random carbon data." *The Journal of Impact and ESG Investing* 4 (2): 8–34. https://doi.org/10.3905/jesg.2023.1.088.

Otchere, I., A. Abdulrahman, and J. Wang. 2025. "Environmental, social, and governance investing and sustainability of pension funds: Evidence from the organisation for economic cooperation and development countries." *Economic Modelling* 143: 106948. https://doi.org/10.1016/j.econmod.2024.106948.

Popescu, I.-S., T. Gibon, C. Hitaj, M. Rubin, and E. Benetto. 2023. "Are SRI funds financing carbon emissions? An input-output life cycle assessment of investment funds." *Ecological Economics* 212: 107918. https://doi.org/10.1016/j.ecolecon.2023.107918.

Riedl, A., and P. Smeets. 2017. "Why do investors hold socially responsible mutual funds?" *The Journal of Finance* 72 (6): 2505–2550. https://doi.org/10.1111/jofi.12547.

Robeco. 2025. *2025 global climate investing survey: Balancing risk, return and sustainability in turbulent times*. https://www.robeco.com/files/docm/docu-202506-robeco-global-climate-investing-survey-2025.pdf.

Setzer, J., and C. Higham. 2025. *Global trends in climate change litigation: 2025 snapshot*. Grantham Research Institute on Climate Change and the Environment, London School of Economics and Political Science. https://doi.org/10.21953/lse.lh46le9y8sgi.

Siemroth, C., and L. Hornuf. 2023. "Why do retail investors pick green investments? A lab-in-the-field experiment with crowdfunders." *Journal of Economic Behavior & Organization* 209: 74–90. https://doi.org/10.1016/j.jebo.2023.02.023.

Sireklove, J. 2016. "Fossil-free investing." *The Journal of Index Investing* 6 (4): 129–133. https://doi.org/10.3905/jii.2016.6.4.129.

STOXX. 2024. *STOXX index methodology guide (Portfolio based indices)*. https://www.stoxx.com/document/News/2024/January/stoxx_index_guide_20240124.pdf.

Task Force on Climate-related Financial Disclosures (TCFD). 2021. *Implementing the recommendations of the Task Force on Climate-related Financial Disclosures.* https://assets.bbhub.io/company/sites/60/2021/07/2021-TCFD-Implementing_Guidance.pdf.

Appendix A. Sector and country weights in STOXX Europe 600 and STOXX Europe TMI, January 2011–December 2024

Panel A: Sector Weights

Sector	STOXX Europe 600	STOXX Europe TMI
Financials	18.9	18.5
Health Care	14.2	13.6
Industrials	13.9	14.5
Consumer Staples	12.4	11.8
Consumer Discretionary	11.4	11.6
Basic Materials	7.1	7.0
Energy	6.9	7.4
Telecommunications	4.8	4.6
Technology	4.7	4.8
Utilities	4.1	4.0
Real Estate	1.7	2.1

Panel B: Country Weights

Country	STOXX Europe 600	STOXX Europe TMI
United Kingdom	27.4	27.3
France	16.0	15.2
Switzerland	14.3	13.9
Germany	13.3	13.1
Netherlands	6.2	7.1
Spain	5.4	5.1
Sweden	4.5	4.5
Italy	4.5	5.1
Denmark	3.8	4.1
Finland	3.1	3.1
Other Countries	2.0	2.1

Notes: Panel A reports average index weights of ICB sectors in the STOXX Europe 600 and STOXX Europe Total Market Index (TMI), expressed as percentages of index free-float market capitalization. Panel B reports the corresponding average country weights. “Other Countries” aggregates all remaining markets with small index weights. Averages are computed over monthly constituent weights from January 2011 to December 2024.

Appendix B. Benchmark Scope 1–2 carbon metrics under alternative estimation methods: STOXX Europe 600, 2011–2024

Refinitiv Datastream provides both reported and model-based estimates of firms' greenhouse gas emissions. In the main analysis, I construct benchmark index-level Scope 1–2 carbon intensity and carbon footprint using only reported emissions and the sector-scaling procedure described in Section 2.2.3.

As robustness, I also consider two alternative treatments of missing firm-level emissions. First, I use Refinitiv's model-based Scope 1–2 estimates for firms that do not report emissions. Second, I implement an author-imputed series based on a Heckman-type selection model that follows partially the approach of Olesiewicz et al. (2023). Figure B.1 compares the resulting benchmark carbon series. The three benchmark series are very similar over time, with only modest differences in the early years when disclosure is sparse: for carbon intensity, the correlations between the reported-only benchmark and the Refinitiv-augmented and author-imputed benchmarks are 0.9996 and 0.9789, respectively, while for the carbon footprint they are 0.9973 and 0.9967. I therefore use the reported-only benchmark in the main analysis, as it is transparent and straightforward to implement.

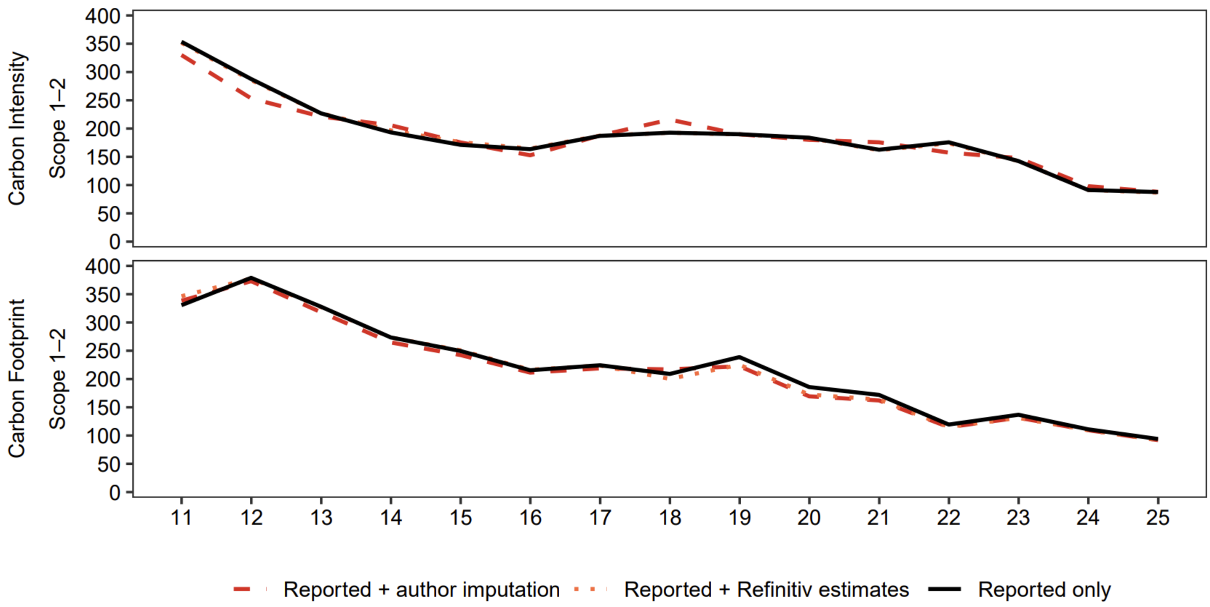


Figure B.1 Alternative estimation methods for carbon measures.

Notes: The figure plots index-level Scope 1–2 carbon intensity (top panel) and carbon footprint (bottom panel) for the STOXX Europe 600 under three alternative treatments of missing firm-level emissions. The solid line uses only reported emissions from Refinitiv. The dotted line combines reported emissions with Refinitiv's model-based Scope 1–2 estimates for non-reporters. The dashed line combines reported emissions with author-imputed values from the Heckman selection model described in Appendix B and based on Olesiewicz, Kooroshy, and Greven (2023). All series are computed using free-float market-value weights. The three methods yield very similar benchmark carbon paths over time, so the main analysis focuses on the reported-only benchmark because it is transparent, easy to replicate, and close to current index-provider practice.

Appendix C. Robustness of main low-carbon portfolio scenarios

Table C.1 Alternative carbon metric: STOXX Europe 600, January 2011–December 2024

Panel A: Objective: Minimize carbon footprint (Scope 1–2) based on market-value

Metric	Benchmark	Conservative	Balanced	Aggressive
ex-ante Tracking Error budget (%)	-	0.50	1.00	2.00
ex-post Tracking Error realized (%)	-	0.70	1.38	2.57
Carbon Footprint Reduction (S1–2 MV, %)	**0.0**	**60.5**	**89**	**96.9**
Carbon Intensity Reduction (S1–2, %)	0.0	50.2	72.8	88.3
Cumulative return (%)	196.1	192.0	190.2	229.1
Annual return (%)	8.06	7.95	7.91	8.88
Annual volatility (%)	13.31	13.36	13.44	13.01
Sharpe Ratio	0.61	0.6	0.6	0.68
Information Ratio	-	-0.16	-0.11	0.32

Panel B: Objective: Minimize carbon footprint (Scope 1–2) based on EVIC

Metric	Benchmark	Conservative	Balanced	Aggressive
ex-ante Tracking Error budget (%)	-	0.50	1.00	2.00
ex-post Tracking Error realized (%)	-	0.67	1.22	2.44
Carbon Footprint Reduction (S1–2 EVIC, %)	**0.0**	**57**	**88**	**96.7**
Carbon Intensity Reduction (S1–2, %)	0.0	49.3	72.7	87
Cumulative return (%)	196.1	193.1	201.3	216.1
Annual return (%)	8.06	7.98	8.2	8.57
Annual volatility (%)	13.31	13.35	13.36	13.04
Sharpe Ratio	0.61	0.6	0.62	0.66
Information Ratio	-	-0.12	0.11	0.21

Notes: The table reports carbon and performance outcomes for three optimized low-carbon portfolios (Conservative, Balanced, Aggressive) and the STOXX Europe 600 benchmark when the optimization objective is Scope 1–2 carbon footprint (bold row). Panel A measures footprint as Scope 1–2 emissions divided by market value, Panel B as Scope 1–2 emissions divided by enterprise value including cash (EVIC). Portfolios use the same tracking-error budgets, constraints, and annual December rebalancing as in Table 3, and all carbon reductions are expressed as percentage reductions relative to the benchmark. Returns and risk statistics are based on monthly total returns in euros, gross of management fees and transaction costs (see Table 3 for metric definitions).

Table C.2 Broader emissions scopes: STOXX Europe 600, January 2011–December 2024

Panel A: Objective: minimize carbon intensity (Scope 1–3)

Metric	Benchmark	Conservative	Balanced	Aggressive
ex-ante Tracking Error budget (%)	-	0.50	1.00	2.00
ex-post Tracking Error realized (%)	-	0.88	1.32	2.26
Carbon Intensity Reduction (S1–3, %)	**0.0**	**29.2**	**66.5**	**94.4**
Carbon Footprint Reduction (S1–3, %)	0.0	21.1	62.9	94.3
Carbon Intensity Reduction (S1–2, %)	0.0	28	62.7	85
Carbon Footprint Reduction (S1–2, %)	0.0	18.9	65.1	88.5
Cumulative return (%)	196.1	199.6	207.5	237.9
Annual return (%)	8.06	8.15	8.35	9.09
Annual volatility (%)	13.31	13.37	13.39	13.59
Sharpe Ratio	0.61	0.62	0.63	0.67
Information Ratio	-	0.1	0.22	0.45

Panel B: Objective: minimize carbon footprint (Scope 1–3)

Metric	Benchmark	Conservative	Balanced	Aggressive
ex-ante Tracking Error budget (%)	-	0.50	1.00	2.00
ex-post Tracking Error realized (%)	-	0.90	1.4	2.14
Carbon Footprint Reduction (S1–3, %)	**0.0**	**25.5**	**73**	**97.8**
Carbon Intensity Reduction (S1–3, %)	0.0	21.4	57.1	83.8
Carbon Intensity Reduction (S1–2, %)	0.0	18.8	50.1	80.7
Carbon Footprint Reduction (S1–2, %)	0.0	23.8	72.5	93.5
Cumulative return (%)	196.1	188.5	186.8	206.9
Annual return (%)	8.06	7.86	7.82	8.34
Annual volatility (%)	13.31	13.24	13.22	13.25
Sharpe Ratio	0.61	0.6	0.6	0.63
Information Ratio	-	-0.22	-0.18	0.13

Notes: This table reports carbon and performance outcomes for three optimized low-carbon portfolios (Conservative, Balanced, Aggressive) and the STOXX Europe 600 benchmark when the optimization objective is based on Scope 1–3 emissions. Panel A minimizes revenue-based Scope 1–3 carbon intensity; Panel B minimizes market-value-based Scope 1–3 carbon footprint. In each case, the table shows reductions in the targeted Scope 1–3 metric and in the corresponding Scope 1–2 metrics, expressed as percentage reductions relative to the benchmark. Portfolios use the same tracking-error budgets, constraint set, and annual December rebalancing as in Table 3. Returns and risk statistics are based on monthly total returns in euros, gross of management fees and transaction costs.

Table C.3 Alternative parent index: STOXX Europe TMI, January 2011–December 2024

Metric	Benchmark	Conservative	Balanced	Aggressive
ex-ante Tracking Error budget (%)	-	0.50	1.00	2.00
ex-post Tracking Error realized (%)	-	0.65	1.24	2.19
Carbon Intensity Reduction (S1–2, %)	**0.0**	**63.5**	**85.6**	**93.9**
Carbon Footprint Reduction (S1–2, %)	0.0	57.7	81.4	93.5
Carbon Intensity Reduction (S1–3, %)	0.0	27.2	50.6	70.2
Carbon Footprint Reduction (S1–3, %)	0.0	19.8	41.5	69.1
Cumulative return (%)	195.6	207.8	234.8	269.5
Annual return (%)	8.05	8.36	9.01	9.79
Annual volatility (%)	13.38	13.38	13.31	13.38
Sharpe Ratio	0.61	0.63	0.68	0.73
Information Ratio	-	0.48	0.78	0.79

Notes: The table reports carbon and performance outcomes for three optimized low-carbon portfolios (Conservative, Balanced, Aggressive) and the STOXX Europe Total Market Index (TMI) benchmark when the objective is to minimize revenue-based Scope 1–2 carbon intensity. Portfolios use the same tracking-error budgets, constraint set, and annual December rebalancing as in Table 3, but are optimized relative to the STOXX Europe TMI rather than the STOXX Europe 600. All carbon reductions are expressed as percentage reductions relative to the TMI benchmark. Returns and risk statistics are based on monthly total returns in euros, gross of management fees and transaction costs (see Table 3 for metric definitions).

Table C.4 Winsorized Scope 1–2 carbon intensity: STOXX Europe 600, January 2011–December 2024

Winsorization level	Benchmark	Conservative	Balanced	Aggressive
0%	0.0	61.0	84.1	93.3
1%	0.0	58.6	82.8	92.7
2%	0.0	56.1	80.9	92
3%	0.0	54.3	79.6	91.4
4%	0.0	52.3	78.3	90.9
5%	0.0	50.8	77.2	90.4

Notes: The table reports Scope 1–2 carbon-intensity reductions for three optimized low-carbon portfolios (Conservative, Balanced, Aggressive) relative to the STOXX Europe 600 benchmark when firm-level revenue-based carbon intensities are winsorized at different two-sided percentile levels within the index universe. For each winsorization level, the optimization objective is to minimize winsorized Scope 1–2 carbon intensity subject to the same tracking-error budgets, constraint set, and annual December rebalancing as in Table 3. Tracking errors are unchanged compared to Table 3 and are omitted for brevity.

Table C.5 Different time windows: STOXX Europe 600

Panel A: Objective: minimize carbon intensity (Scope 1–2), January 2016 - December 2024

Metric	Benchmark	Conservative	Balanced	Aggressive
ex-ante Tracking Error budget (%)	-	0.50	1.00	2.00
ex-post Tracking Error realized (%)	-	0.67	1.22	2.22
Carbon Intensity Reduction (S1–2, %)	**0.0**	**62.5**	**86.1**	**94.4**
Carbon Footprint Reduction (S1–2, %)	0.0	61	81.2	93.4
Cumulative return (%)	87.2	90.4	96.1	105.1
Annual return (%)	7.21	7.42	7.77	8.31
Annual volatility (%)	13.56	13.48	13.42	13.56
Sharpe Ratio	0.54	0.56	0.59	0.62
Information Ratio	-	0.31	0.46	0.49

Panel B: Objective: minimize carbon intensity (Scope 1–2), January 2020 - December 2024

Metric	Benchmark	Conservative	Balanced	Aggressive
ex-ante Tracking Error budget (%)	-	0.50	1.00	2.00
ex-post Tracking Error realized (%)	-	0.73	1.28	2.36
Carbon Intensity Reduction (S1–2, %)	**0.0**	**60.3**	**87.7**	**95.2**
Carbon Footprint Reduction (S1–2, %)	0.0	61.8	84.9	94.8
Cumulative return (%)	43.2	46.7	48.6	47.9
Annual return (%)	7.44	7.96	8.24	8.14
Annual volatility (%)	15.71	15.52	15.44	15.54
Sharpe Ratio	0.5	0.54	0.56	0.55
Information Ratio	-	0.71	0.62	0.3

Notes: This table reports carbon and performance outcomes for three optimized low-carbon portfolios (Conservative, Balanced, Aggressive) and the STOXX Europe 600 benchmark over two shorter analysis windows. In both panels, portfolios minimize revenue-based Scope 1–2 carbon intensity using the same tracking-error budgets, constraint set, and annual December rebalancing as in Table 3. Panel A covers the post-Paris period 2016–2024; Panel B focuses on the recent years 2020–2024. Carbon-intensity and carbon-footprint reductions are expressed as percentage reductions relative to the benchmark. Returns and risk statistics are based on monthly total returns in euros, gross of management fees and transaction costs (see Table 3 for metric definitions).

Table C.6 Implementation, sustainability, and factor exposures: STOXX Europe 600, January 2011–December 2024

Panel A: Implementation and trading characteristics

Metric	Benchmark	Conservative	Balanced	Aggressive
Cumulative return, gross (%)	196.1	209.2	230.8	267.5
Cumulative return, net (%)	-	196.9	216.9	252.4
Annual return, gross (%)	8.06	8.40	8.92	9.74
Annual return, net (%)	-	8.08	8.59	9.41
Information Ratio (gross)	-	0.52	0.73	0.81
Information Ratio (net)	-	0.03	0.45	0.65
Average active weight (%)	-	21.5	37.4	62.1
Average turnover (%)	-	44.9	53.3	51.5
Average number of stocks	596	406	292	156

Panel B: Sustainability performance

Sustainability Score	Benchmark	Conservative	Balanced	Aggressive
Refinitiv (LSEG)				
Refinitiv ESG Combined	62.5	63.2	63.9	64.3
Refinitiv ESG	73.9	74.7	74.6	73.1
Refinitiv Environmental pillar	75.8	77.2	77.0	74.8
Refinitiv Resource use Score	81.9	83.9	84.4	83.2
MSCI				
MSCI ESG	6.8	6.9	7.0	7.1
MSCI Environmental pillar	6.4	6.5	6.6	6.9

Panel C: (Fama-French) Factor Exposure

Factor	Conservative	Balanced	Aggressive
Market	0.00	0.00	0.01
Small Minus Big (SMB)	-0.01	-0.01	0.03
High Minus Low (HML)	0.00	-0.04**	-0.11***
Robust Minus Weak (RMW)	0.00	-0.01	-0.06
Conservative Minus Aggressive (CMA)	0.01	0.04*	0.04
Momentum (MOM)	0.01	0.02	0.04**

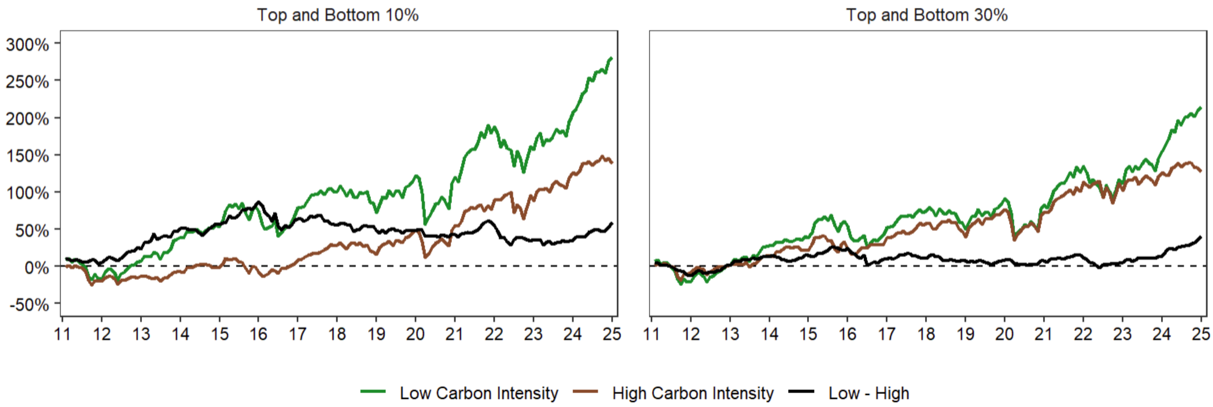
Notes: The table reports implementation, sustainability, and factor-exposure characteristics for three optimized low-carbon portfolios (Conservative, Balanced, Aggressive) and the STOXX Europe 600 benchmark. In all panels, portfolios are optimized to minimize revenue-based Scope 1–2 carbon intensity using the same tracking-error budgets, constraint set, and annual December rebalancing as in Table 3. Panel A shows implementation and trading characteristics. Cumulative and annual returns are reported gross and net of an annual management fee and proportional transaction costs applied to turnover at each rebalance (see Section 3.4.2 for details). Average active weight is the mean sum of absolute benchmark-relative weights, and turnover is the average one-way annual turnover of the optimized portfolios. The average number of stocks is the mean number of holdings over the sample. Panel B shows sustainability performance based on Refinitiv (LSEG) and MSCI scores. Refinitiv ESG Combined, ESG, Environmental pillar, and Resource Use scores are provider-defined measures on a 0–100 scale, with higher values indicating better sustainability performance. MSCI ESG and Environmental pillar scores are reported on a 0–10 scale, with higher values indicating better performance. Panel C shows Fama–French factor exposures. Entries report differences in factor loadings $\Delta\beta = \beta_{Portfolio} - \beta_{Benchmark}$ from time-series regressions of monthly portfolio and benchmark returns on the Fama–French six-factor model (Market, SMB, HML, RMW, CMA, and Momentum). Asterisks indicate statistical significance of $\Delta\beta$ at the 10% (*), 5% (**), and 1% (***) levels.

Table C.7 US investor perspective: STOXX Europe 600 low-carbon portfolios in USD, January 2011–December 2024

Metric	Benchmark	Conservative	Balanced	Aggressive
ex-ante Tracking Error budget (%)	-	0.50	1.00	2.00
ex-post Tracking Error realized (%)	-	0.63	1.17	2.09
Carbon Intensity Reduction (S1–2, %)	**0.0**	**61.1**	**84.1**	**93.3**
Carbon Footprint Reduction (S1–2, %)	0.0	57.5	79.4	93.3
Cumulative return (%)	127.1	136.7	152.4	180.6
Annual return (%)	6.03	6.35	6.84	7.65
Annual volatility (%)	16.61	16.56	16.49	16.56
Sharpe Ratio	0.4	0.42	0.45	0.5
Information Ratio	-	0.5	0.69	0.77

Notes: The table reports tracking-error budgets, carbon reductions, and performance statistics for three optimized low-carbon portfolios (Conservative, Balanced, Aggressive) and the STOXX Europe 600 benchmark from the perspective of a US-based investor. Portfolios are identical to the main scenarios in Table 3 and minimize revenue-based Scope 1–2 carbon intensity subject to the same tracking-error budgets, constraint set, and annual December rebalancing; only the reporting currency is changed from euros to US dollars. Carbon-intensity and carbon-footprint reductions for Scope 1–2 are expressed as percentage reductions relative to the benchmark. Cumulative and annual returns, volatility, Sharpe ratios, and information ratios are based on monthly total returns converted to USD, gross of management fees and transaction costs.

Appendix D. Performance of carbon-intensity-sorted portfolios



Notes: This figure plots cumulative returns for value-weighted portfolios formed on monthly lagged carbon intensity (CI). Stocks are sorted into top and bottom 10% and 30% by Scope 1–2 CI. Portfolios are rebalanced monthly. The black line shows the cumulative return spread (Low CI minus High CI).

Appendix E. Variables and Refinitiv Codes

Panel A: Firm-level financial and classification variables

Variable	Code
Monthly stock return	RI
Market value (MV)	MV
Free-Float Market value (MVFF)	MVFF
Revenue (REV)	WC01001
Common equity (CE)	WC03501
Total Assets (TA)	WC02999
Earnings before Interest and Taxes (EBIT)	WC18191
Enterprise Value (EV)*	WC18100
Cash*	WC02003
Cash & Due from Banks*	WC02004
Cash & Equivalents Generic*	WC02005
ICB Sector Name*	ICBSN
ICB Industry Name (Sector)	ICBIN
Geographical Classification of Company	GEOGN

* Used to construct EVIC.

Panel B: Firm-level emission variables

Variable	Short description	Code
Scope 1 Emissions	Direct CO_2 and CO_2 equivalents emission in tonnes from that are owned or controlled by the company.	ENERDP024
Scope 2 Emissions	Indirect CO_2 and CO_2 equivalents emission in tonnes from consumption of purchased electricity, heat or steam which occur at the facility where electricity, steam or heat is generated.	ENERDP025
Scope 1–2 Emissions	Note: I use this data to verify the Scope 1 (Refinitiv) and Scope 2 (Refinitiv) data points.	ENERDP023
Scope 1–2 Emissions (Estimated)	Note: I use this data in Appendix B to compare alternative treatments of missing firm-level emissions to calculate the index's carbon intensity.	ENERDP123
Scope 3 Emissions	Total CO_2 and CO_2 Scope Three equivalent emission in tonnes. Scope 3 includes: (i) emissions from contractor-owned vehicles, employee business travel (by rail or air), waste disposal, outsourced activities, and (ii) emissions from product use by customers, emission from the production of purchased materials, emissions from electricity purchased for resale.	ENERDP096

Online Appendix A. Results of related literature

Online Appendix Table A.1. Summary of decarbonization outcomes in related equity studies

Author	Parent Index	Horizon	Method	Scopes	Carbon Reduction (%)		Tracking Error (%)		Return (ann.) (%)	
(1)	(2)	(3)	(4)	(5)	CI (6)	CF (7)	ex ante (8)	ex post (9)	Port (10)	BM (11)
Jondeau et al. (2025)[a1]	MSCI ACWI	2005 – 2020	Exclusion	1–3	68.0	70.7	-	1.71	10.15	9.69
Jondeau et al. (2025) [a2]	MSCI ACWI	2005 – 2020	Best-in-class	1–3	61.9	62.8	-	1.06	9.72	9.69
Jondeau et al. (2025)[a3]	MSCI ACWI	2005 – 2020	Dynamic Best-in-class	1–3	64.2	-	-	1.05	9.56	9.69
Jondeau et al. (2025)[a4]	MSCI ACWI (Europe)	2005 – 2020	Dynamic Best-in-class	1–3	71.1	-	-	1.88	7.62	7.55
Jondeau et al. (2025)[a5]	MSCI ACWI (Europe)	2005 – 2020	Dynamic Best-in-class	1–3	74.3	-	-	2.79	7.49	7.55
Bouchey et al. (2024)[b1]	S&P DM (LM)	2010 – 2022	Screening	1–3	57.6	53.8	-	1.01	9.44	8.93
Bouchey et al. (2024)[b2]	S&P DM (LM)	2010 – 2022	Optimization	1–3	53.1	58.4	1.00	1.14	9.24	8.93
Bouchey et al. (2024)[b3]	S&P DM (LM)	2010 – 2022	Optimization	1–3	66.2	-	-	1.66	-	-
Blitz et al. (2024)[c1]	MSCI World	1989 – 2022	Optimization	1–2	-	59.2	0.50	0.53	-	-
Blitz et al. (2024)[c2]	MSCI World	1989 – 2022	Optimization	1–2	-	73.2	1.00	0.98	-	-
Fahlenbrach and Jondeau (2023)[d]	SNB (MSCI USA)	2014 – 2021	Exclusion	1–2	61.2	53.4	-	0.23 (1.19)	15.4	15.5
Bender et al. (2019)[e1]	MSCI World	2013 – 2018	Optimization	1–3	64.1	-	1.00	1.43	11.29	9.94
Bender et al. (2019)[e2]	MSCI World	2013 – 2018	Optimization	1–3	72.1	-	2.00	2.58	11.34	9.94
Bender et al. (2019)[e3]	MSCI World	2013 – 2018	Optimization	1–3	72.5	-	3.00	3.15	12.67	9.94
Andersson et al. (2016)[f1]	MSCI Europe	2010 – 2016	Index	1–2	52.0	-	-	0.70	8.70	7.80
Andersson et al. (2016)[f2]	MSCI Europe	2010 – 2014	Index-rules	1–2	62.0	-	-	0.72	-	-
Andersson et al. (2016)[f3]	MSCI Europe	2010 – 2014	Optimization	1–2	82.0	-	-	0.90	-	-

Notes: The table summarizes key design choices and outcomes in selected studies on equity portfolio decarbonization. Columns (1)–(5) report the parent index, sample horizon, portfolio-construction method, and emissions scopes considered. Columns (6) and (7) show percentage reductions in weighted-average carbon intensity (CI) and carbon footprint (CF) relative to the parent index, where reported. Columns (8) and (9) report ex ante and ex post tracking error in percent per year; columns (10) and (11) show annualized returns of the decarbonized portfolio (Port) and the benchmark (BM). A dash indicates that the value is not reported. Where noted, I compute CI or CF reductions from reported WACI or footprint values in the original tables. See next page for source notes.

Online Appendix Table A.1. (continued). Source Notes:

[a] Jondeau et al. (2025) examine the MSCI All Country World Index (ACWI) from 2005 to 2020 with the following carbon reduction strategies: Exclusion[a1] (p. 12, Table 2, 75%), best-in-class[a2] (p. 17, Table 4, 75%), Dynamic best-in-class[a3] (p. 20, Table 5, PAB), and Regional Dynamic best-in-class (p. 24, Table 6, Europe, 10%[a4] & PAB[a5]). For the dynamic best-in-class, I self-calculate the average carbon reduction from their tables by dividing the respective portfolio value by its benchmark value. Further, I use their WACI rows for CI, as they are defined as our CI.
[b] Bouchey et al. (2024) analyze the S&P Developed Market (DM) Index including large- and mid-caps (LM) from 2010 to 2022. Values are taken from Exhibit 3 (p. 63) for Screening[b1], case B) and for Optimization[b2], case F). The second Optimization[b3] is taken from Exhibit 7, row Sectors & Industries ±20% without TE constraint (p. 67).
[c] Blitz et al. (2024) investigate the MSCI World from 1989 to 2022 with a 3D optimization approach, that jointly optimizes expected return, risk, and sustainability. Values are taken from their Table 2 (p. 10). Their carbon footprint is calculated with EVIC instead of MV as denominator.
[d] Fahlenbrach and Jondeau (2023) decarbonize the Swiss National Bank's (SNB) U.S. equity portfolio (banks excluded). The reported CI and CF reductions are measured relative to the initial SNB portfolio (averaged over 2013–2020). Values are taken from Table 4, Panel C, 5% (p. 816).
[e] Bender et al. (2019) investigate the MSCI World from July 2013 to June 2018. I extract the portfolio performance metrics from Exhibit 10 (p. 15) and the carbon reduction metrics from Exhibit 11 (p. 15).
[f] Andersson et al. (2016) analyze decarbonized versions of the MSCI Europe Index. In Table 2 (p. 23) they report the characteristics and performance of the MSCI Europe Low Carbon (LC) Leaders Index, constructed with transparent screening / index rules, relative to its parent index MSCI Europe over November 2010–February 2016 (backtest). In Table 3 (p. 23) they compare two MSCI-Europe-based decarbonized portfolios over 30 November 2010–30 June 2014: an optimized low-carbon target portfolio and a transparent rules-based portfolio. Tracking Errors are always measured relative to MSCI Europe.

Online Appendix B. European equity risk model

This appendix describes in detail how I construct the multi-factor European equity risk model that creates the covariance matrix used to calculate the ex-ante tracking-error from equation (7). The risk model is a linear factor model with style, market, country and sector factors, and is based on the linear multi-factor equity risk-model framework described in Grinold and Kahn (2000, chapter 3). For convenience, I provide the relevant Datastream (Worldscope) and Refinitiv codes in italics in brackets, as well as the key functions used from R packages where helpful.

B.1 Style factors, country exposures, and sector exposures

My universe of potential stocks at each month-end t consists of all constituents of the STOXX Europe 600 (*LDJSTOXX*) and STOXX Europe TMI (*LDJTMSTE*).[1] All accounting variables were lagged by one year to avoid look-ahead bias. At each month-end t, I compute six firm-level characteristics for every stock i in the universe:

1. Size is given by the logged free-float market value (*MVFF*):

$$SIZE_{i,t} = \log\left(MVFF_{i,t}\right) \qquad (B.1)$$

2. Value is defined as the logged ratio of lagged common equity (*WC03501*) to MVFF, for stocks with a positive common equity and positive MVFF:

$$BM_{i,t} = \frac{Common\ Equity_{i,t-1}}{MVFF_{i,t}} \qquad (B.2)$$

$$VALUE_{i,t} = \log\left(BM_{i,t}\right) \qquad (B.3)$$

3. Quality is measured as EBIT (*WC18191*) divided by Total Assets (*WC02999*), both lagged by one year:

$$QUALITY_{i,t} = \frac{EBIT_{i,t-1}}{Total\ Assets_{i,t-1}} \qquad (B.4)$$

4. Investment captures the growth in total assets over the prior year:

$$INVESTMENT_{i,t} = \frac{Total\ Assets_{i,t-1}}{Total\ Assets_{i,t-2}} - 1 \qquad (B.5)$$

[1] To obtain the constituents for a specific month, append the month (MM) and year (YY) to the list code. For example, to retrieve the STOXX Europe 600 constituents for December 2024, use: *LDJSTOXX1224*.

5. Momentum is defined as the cumulative total return over the past 12 months excluding the most recent month. Total returns are computed from the Datastream total return index (*RI*). Let $r_{i,\tau}$ denote the monthly total stock return of stock i in month τ, then:

$$MOM_{i,t} = \prod_{\tau=t-12}^{t-2} \left(1 + r_{i,\tau}\right) - 1 \qquad (B.6)$$

6. Volatility is the standard deviation of monthly total returns over the past 12 months:

$$VOL_{i,t} = sd\left(r_{i,t-12}, \dots, r_{i,t}\right) \qquad (B.7)$$

Within each month t, I winsorize each style characteristic at the 1st and 99th cross-sectional percentiles and then standardize it to a z-score.

For the country and sector exposures, I use dummy variables constructed from Refinitiv's geographical classification of a company (GEOGN) and the ICB industry classification (ICBIN). I form one country dummy for each of the ten largest markets in the STOXX universe (United Kingdom, France, Switzerland, Germany, Netherlands, Spain, Sweden, Italy, Denmark, and Finland), and group all remaining countries into an "Other" bucket. Each dummy equals one if stock i belongs to that country at month-end t, and zero otherwise. Sector exposures are defined analogously using the 11 ICB industries in my sample (Financials, Telecommunications, Energy, Industrials, Basic Materials, Real Estate, Consumer Staples, Utilities, Health Care, Consumer Discretionary, Technology). Finally, the market factor exposure is set to 1 for all stocks at all dates.

B.2 Factor returns and factor covariance

Factor returns are estimated monthly using the STOXX Europe 600 and STOXX Europe TMI constituents. For each month t, let $r_{i,t}$ denote the total return of stock i, and let $Z_{i,t}$ denote the corresponding vector of standardized (z-scored) style characteristics from section B.1.

For the style factors, I estimate factor returns via a value-weighted cross-sectional regression. In each month t, I calculate

$$r_{i,t} = \alpha_t + \sum_{k \in K^{style}} f_{k,t}^{style} Z_{i,k,t} + \varepsilon_{i,t} \qquad (B.8)$$

where $K^{style} = \{\text{Size, Value, Quality, Investment, MOM, VOL}\}$, and the regression is estimated by weighted least squares with weights given by the stock's free-float market value $MVFF_{i,t}$. The betas of $f_{k,t}^{style}$ are interpreted as the monthly style factor returns.

For the market f_t^{market}, country $f_t^{country}$, and sector f_t^{sector} factors, I compute value-weighted portfolio returns. The market factor is the value-weighted return of all STOXX Europe TMI constituents. Country factors are value-weighted returns within each of the ten country groups defined in Section B.1 (plus "Other"). Sector factors are defined analogously for each of the 11 ICB industries.

By stacking all factors, I obtain a vector of factor return f_t containing the six style factor returns, the market return, country returns and sector returns

$$f_t = \left(f_t^{style}, f_t^{market}, f_t^{country}, f_t^{sector}\right)^\top \qquad (B.9)$$

The time-varying factor covariance matrix $\Sigma_{f,t}$, is computed from the time series of monthly factor returns using an exponentially weighted moving average (EWMA) with decay parameter λ=0.97. The EWMA weight on month τ when estimating at time t is

$$\omega_{t,\tau} = \frac{\lambda^{t-\tau}}{\sum_{j=0}^{t-1} \lambda^j}, \quad \tau = 1, \dots, t \qquad (B.10)$$

So that $\omega_{t,\tau}$ declines geometrically as τ moves further into the past and $\sum_{\tau=0}^{t-1} \omega_{t,\tau} = 1$.

The EWMA mean of factor returns at time t is then

$$\overline{f_t} = \sum_{\tau=1}^{t} \omega_{t,\tau} f_\tau \qquad (B.11)$$

and the corresponding EWMA factor covariance matrix $\Sigma_{f,t}$is

$$\Sigma_{f,t} = \sum_{\tau=1}^{t} \omega_{t,\tau} \left(f_\tau - \overline{f_t}\right)\left(f_\tau - \overline{f_t}\right)^\top \qquad (B.12)$$

B.3 Specific risk

The specific risk of stock i captures the part of its returns not explained by the common factors. For each month t, I use the residuals $\varepsilon_{i,t}$ from regression (B.8) to build the time-varying specific variance with an EWMA recursion:

$$\sigma_{\epsilon,i,t}^2 = \lambda \sigma_{\epsilon,i,t-1}^2 + (1-\lambda)\epsilon_{i,t}^2 \text{ with } \lambda = 0.97 \qquad (B.13)$$

If a stock is missing residuals, the previous period's residuals are carried forward. Any remaining missing values are imputed with the cross-sectional median. This results in a diagonal matrix of specific variances

$$D_t = diag\left(\sigma^2_{\epsilon,1,t}, \dots, \sigma^2_{\epsilon,N,t}\right) \quad (B.14)$$

where N_t is the number of stocks in the respective universe at month end t.

B.4 Total covariance matrix and tracking error

The total covariance matrix is obtained by combining the common-factor risk and the specific risk. Let B_t denote the $N_t \times K$ matrix of factor exposures at month-end t, where each row corresponds to one of the N_t stocks. The k-th column of B_t contains the exposures of all stocks to factor k (style, market, country, or sector).

Given the factor covariance matrix $\Sigma_{f,t}$ from equation (B.12) and the diagonal matrix of specific variances D_t from equation (B.14), the model-implied covariance matrix of stock returns at month-end t is

$$\Sigma_{\mathrm{t}} = B_t \Sigma_{f,t} B_t^{\top} + D_t \quad (B.15)$$

In practice, numeric imperfections can result in a Σ_{t} that is not strictly positive semi-definite. To address this, I apply a near-positive-definite adjustment (using Matrix::nearPD in R) and add a small ridge term to the diagonal, scaled proportionally to the average variance, which ensures numerical stability in the portfolio optimization.

The resulting Σ_{t} is then used to compute the ex-ante tracking-error in equation (7) for each optimized portfolio at each annual rebalancing date in December:

$$TE_t\left(\omega_{P,t}\right) = \sqrt{\left(\omega_{P,t} - \omega_{I,t}\right)^{\top} \Sigma_t \left(\omega_{P,t} - \omega_{I,t}\right)} \quad (B.16)$$

where $\omega_{P,t}$ and $\omega_{I,t}$ are the portfolio and index weights and Σ_t is the covariance matrix implied by my risk model.

Paper 4 – 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

* This paper is joint work with Huong Dang and Manuel Brinkmann. It is currently being revised for the next submission. A working paper version is available as: Heldmann, J., H. D. Dang, and M. Brinkmann. 2025. "Sustainability ratings, equity portfolio performance, and factor models: Evidence from a multi-specification approach." Presented at FoFI 2026. https://wp.lancs.ac.uk/fofi2026/files/2026/03/FoFI-2026-027-Jan-Heldmann_F.pdf. This version includes minor editorial adjustments.

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] Dobrick et al.

[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).

(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; Soebhag 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.

[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.

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 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).

[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.

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 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, 1989) 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-

[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.

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 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* 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

[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.

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.

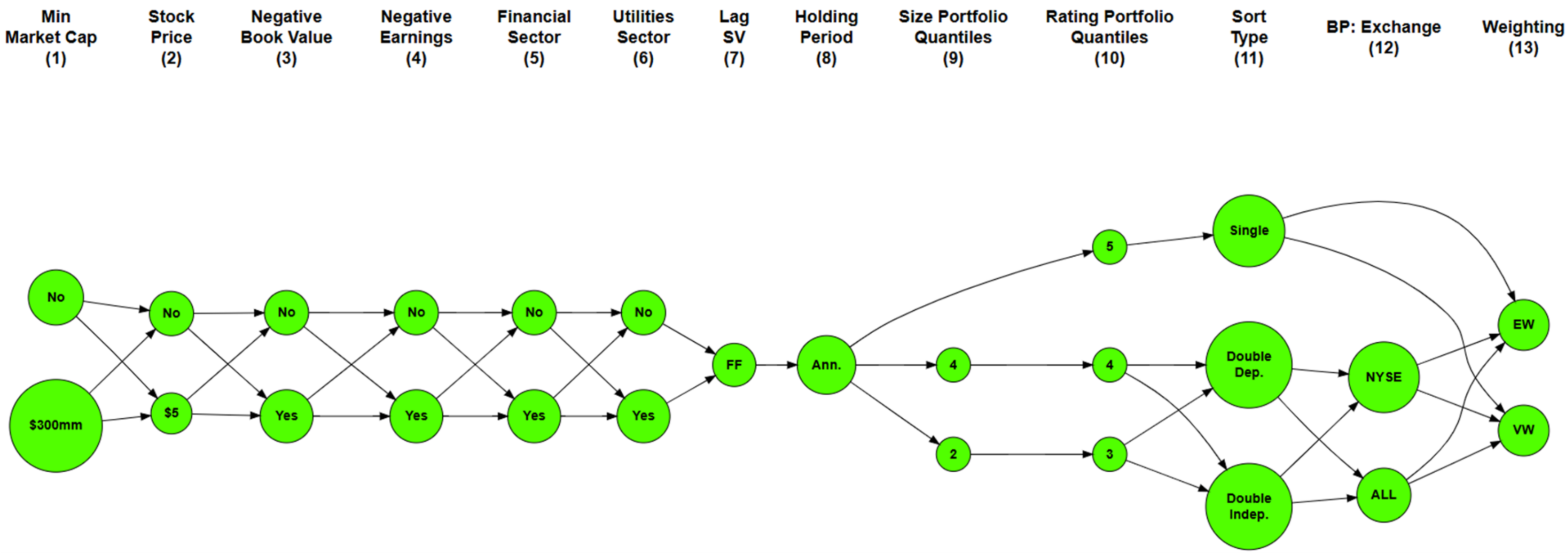
Figure 1 is a flowchart featuring the decisions we consider at each step (node) in constructing our rating-sorted portfolios.[11] The first decision (1) relates to the exclusion of companies based on their market capitalization. Small or very small companies make up a large part of the U.S. equity market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

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

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 quantiles based on rating scores (5x1, 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.

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 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.

---

[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 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.

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 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-

[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.

[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.

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 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

[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.

[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.

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 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.

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).

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

[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.

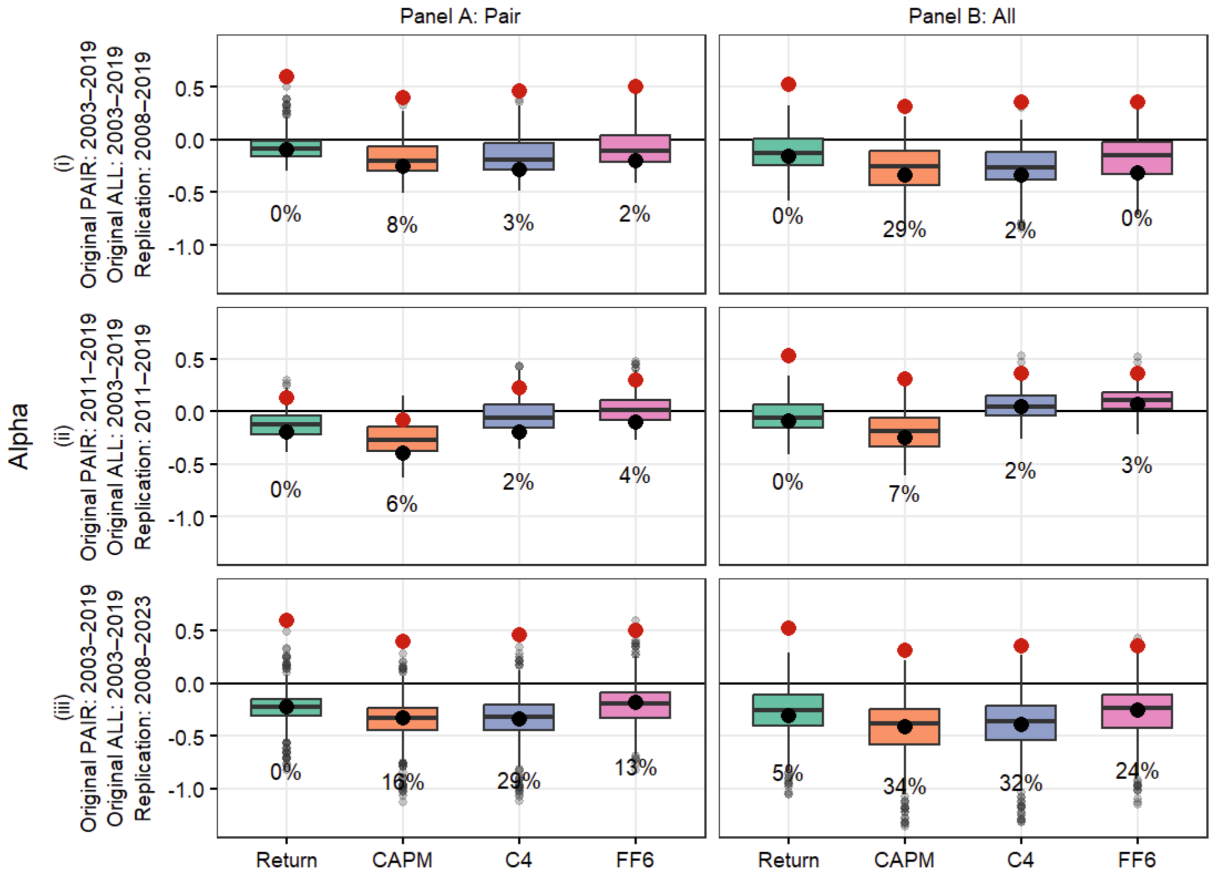


Figure 2. Avramov et al. (2022) Replication: Low ESG Uncertainty, Low ESG Rating minus Low ESG Uncertainty, High ESG Rating return (Low LMH-R return)

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.

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 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.

[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.

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 Returns 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 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

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

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)

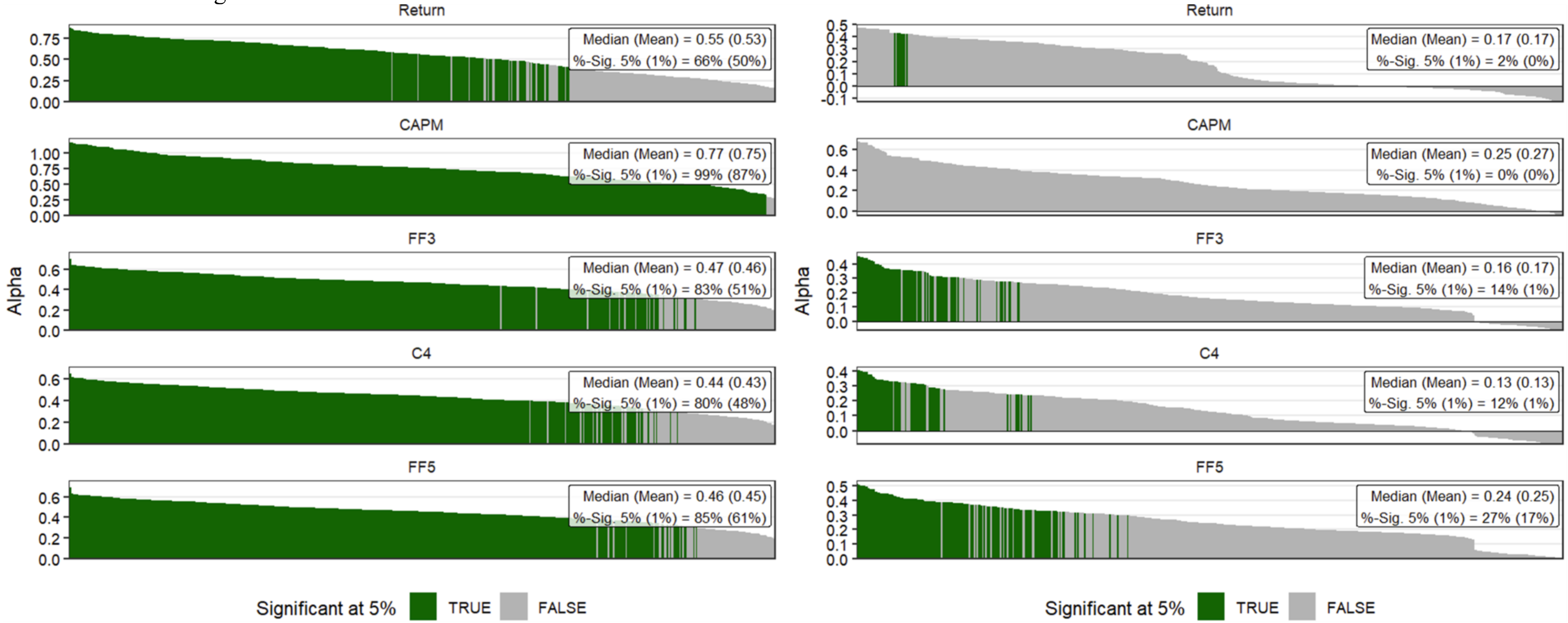
We replicate Pástor et al.'s (2022) analysis that examines whether the GMB spreads can be well explained by factors in traditional asset pricing models (Table 3 in their article). Specifically, we repeat the analysis corresponding to models in columns 1 (Return), 2 (CAPM), 3 (Fama-French three-factor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

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

[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.

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.

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

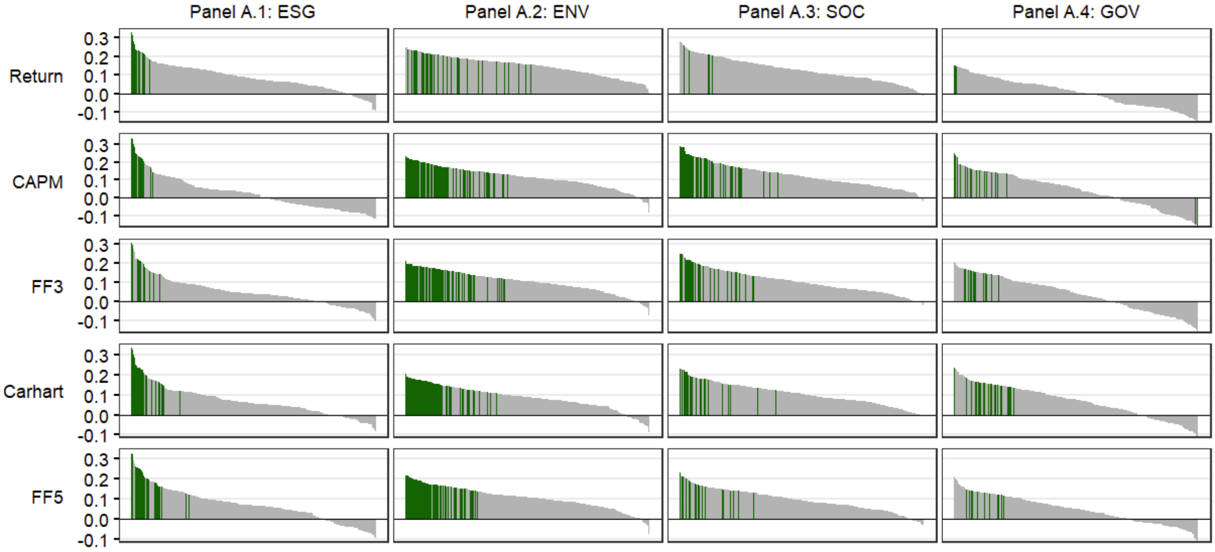
[33] As indicated in Online Appendix C.1., restricted by data availability, our sample does not include ratings from MSCI KLD, which were last available in 2018, and the two less popular raters, namely, FTSE and Inrate.

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).

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.

Panel A. Original study period 2010–2017



Panel B. Extended study period 2010–2023

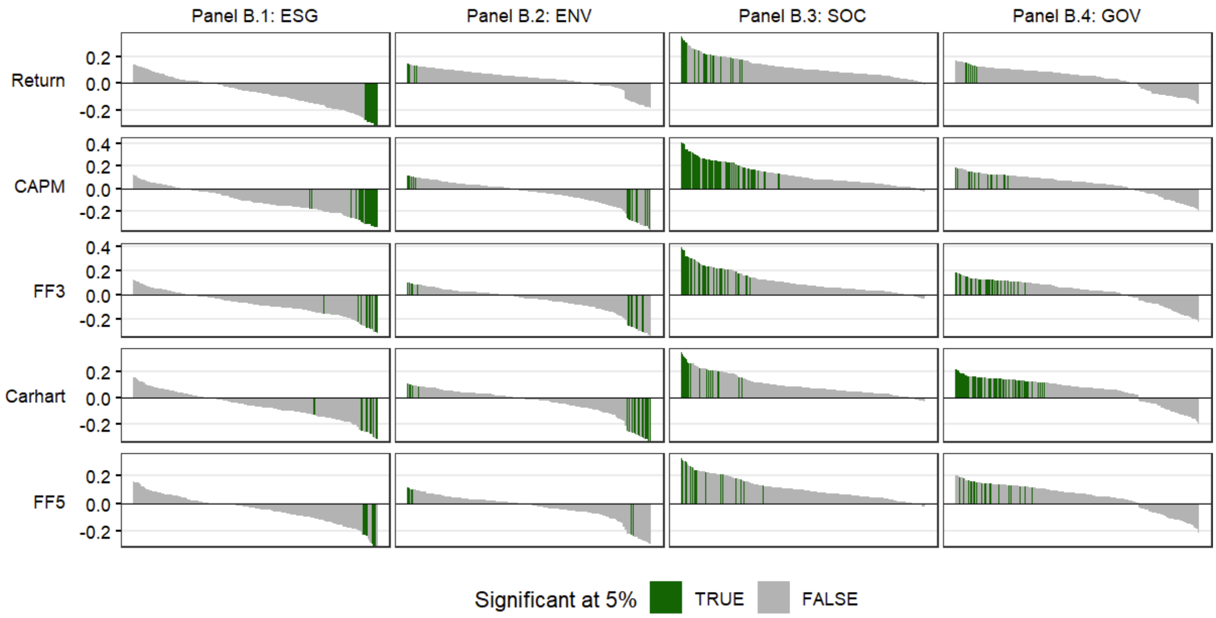


Figure 4. Gibson-Brandon et al. (2021) Table 5 Replication – Alpha estimate distribution using our multi-specification approach

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 2010-2023, are insignificant.

4 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.

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:

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

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

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]

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

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.

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.

[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.

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). 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

[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.

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 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.

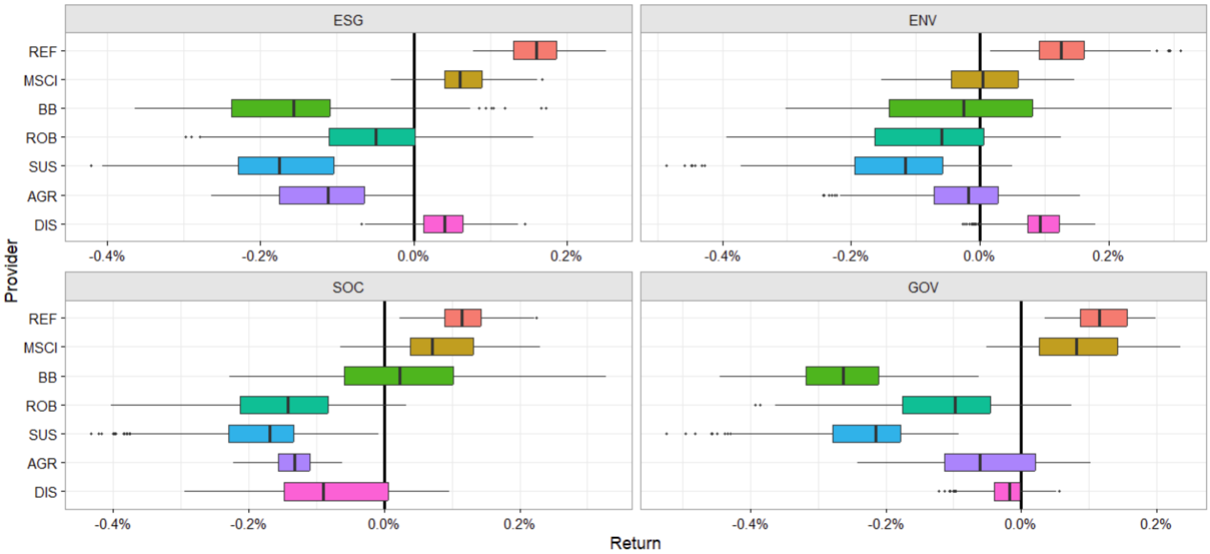
To examine whether rating factors generate positive (risk-adjusted) returns, we first examine their returns and Sharpe ratios using box plots. If companies with high rating scores perform better financially than companies with lower scores, we should expect a box in the boxplot to be in the positive 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).

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).

Panel A: Factor Returns



Panel B: Sharpe Ratios

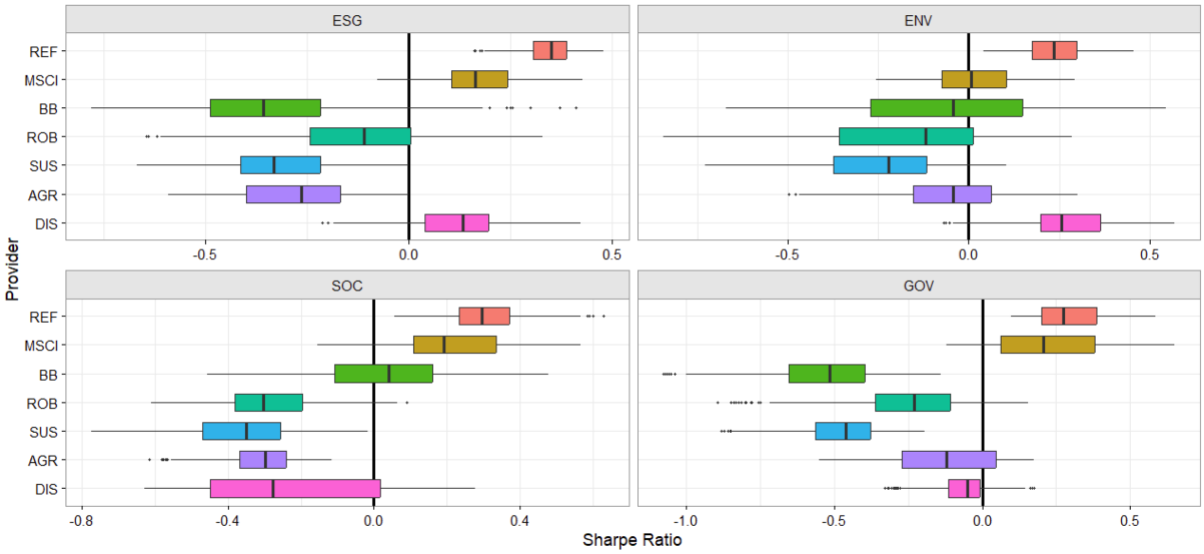


Figure 5. Rating factors: excess returns and Sharpe ratios

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.

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, ΔaSR2) between the extended portfolio and the base (FF6) portfolio.

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.

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.

[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.

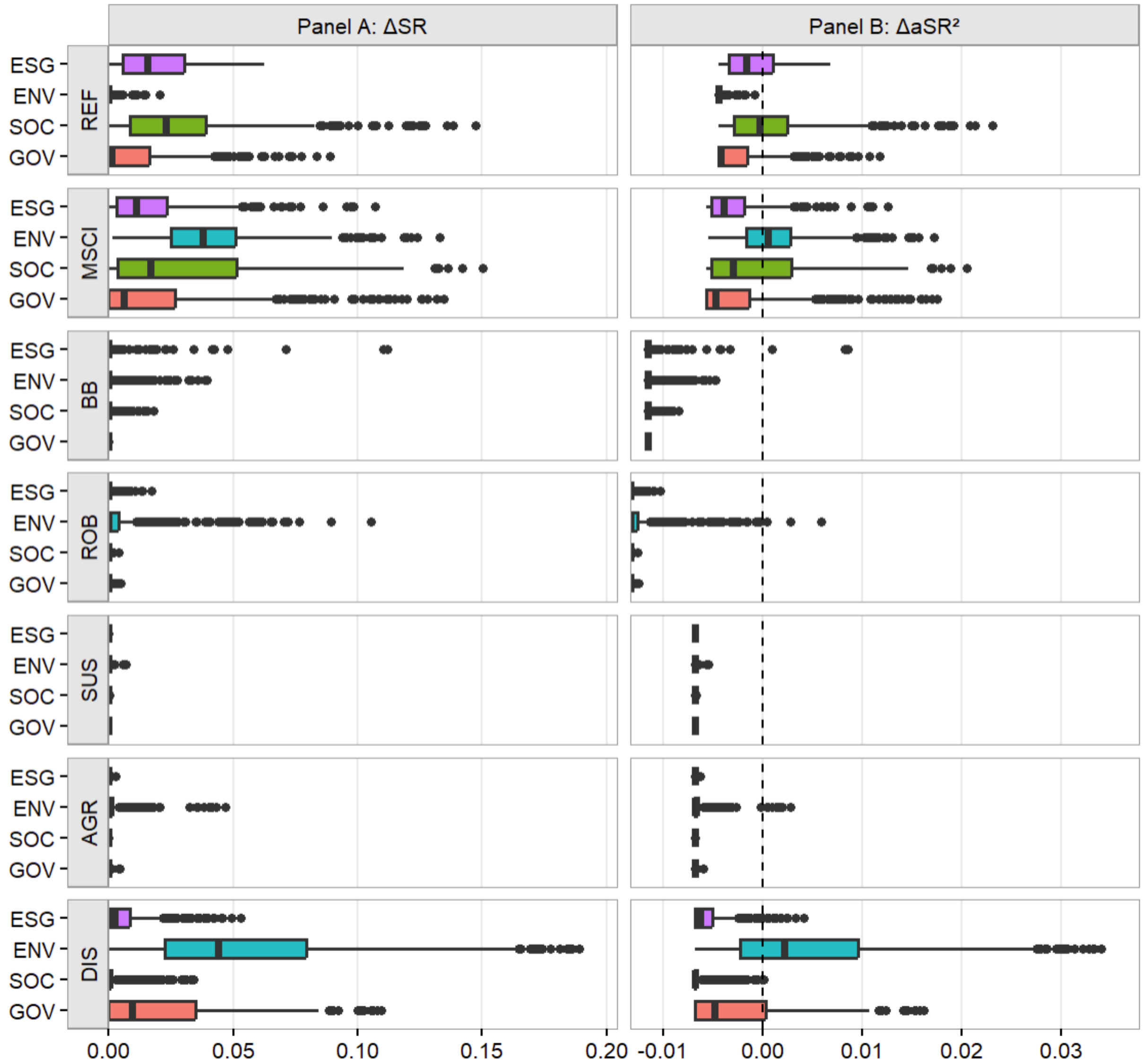


Figure 6. Ex-post max Sharpe ratios

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.

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:

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

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

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).

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).

Table 1. Rating factor spanning

Provider, analysis period (1)	Rating (2)	RM (3)	SMB5 (4)	HML (5)	RMW (6)	CMA (7)	MOM (8)	R^2 (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 (1)	Rating (2)	RM (3)	SMB5 (4)	HML (5)	RMW (6)	CMA (7)	MOM (8)	R^2 (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.

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:

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

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_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

[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.

[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.

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.

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.

[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.

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)						
(1)	%Δ (Δ) p(GRS-Test) <5% (2)	Δ $\overline{R^2}$ (3)	Δ $\overline{R^2_H - R^2_L}$ (4)	%Δ (Δ) $p(\alpha) < 1\%$ (5)	%Δ (Δ) $p(\alpha) < 5\%$ (6)	Δ $\overline{\lvert\alpha\rvert}$ (7)	Δ $\overline{\alpha_H - \alpha_L}$ (8)	Δ p(GRS-Test) <5% (9)	Δ $\overline{R^2}$ (10)	Δ $\overline{R^2_H - R^2_L}$ (11)	%Δ (Δ) $p(\alpha) < 1\%$ (12)	%Δ (Δ) $p(\alpha) < 5\%$ (13)	Δ $\overline{\lvert\alpha\rvert}$ (14)	Δ $\overline{\alpha_H - \alpha_L}$ (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$\overline{\lvert\alpha\rvert}$ (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$\overline{\lvert\alpha\rvert}$, 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

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

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

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.

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, Part B): 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, December 10, 2024. Accessed 20 July, 2025. https://web.archive.org/web/20250725165911/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.

Cakici, Nusret, Christian Fieberg, Gábor Neszveda, Vanja Piljak, and Adam Zaremba. 2025. "Lost in the Multiverse: Methodological Uncertainty in Studying Global Equity Returns." SSRN Working Paper. https://doi.org/10.2139/ssrn.5181455.

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.

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. https://doi.org/10.1201/9781003215257.

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.

Gibbons, Michael. R., Stephen. A. Ross, and Jay. Shanken. 1989. "A Test of the Efficiency of a Given Portfolio." *Econometrica* 57 (5): 1121–1152. https://doi.org/10.2307/1913625.

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 Working Paper. https://doi.org/10.2139/ssrn.3546295.

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 US 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. https://www.abacademies.org/articles/exploratory-review-of-esg-factor-attribution-to-the-portfolio-return-in-famafrench-factor-model-framework-15711.html.

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 Working Paper. 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.

Soebhag, Amar, Bart Van Vliet, and Patrick Verwijmeren. 2024. "Non-standard Errors in Asset Pricing: Mind your Sorts." *Journal of Empirical Finance* 78: 101517. https://doi.org/10.1016/j.jempfin.2024.101517.

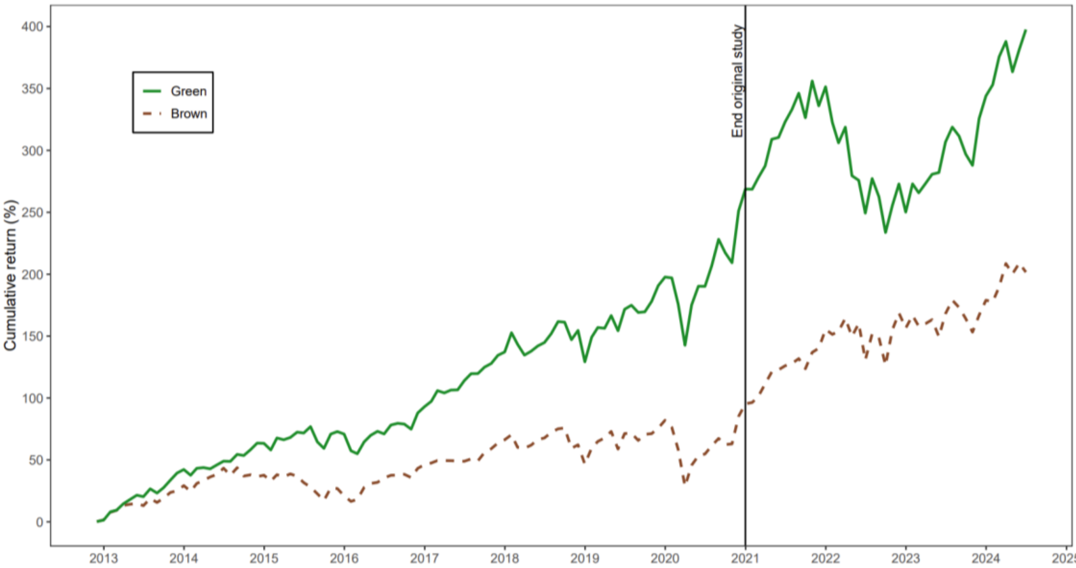
Walter, Dominik, Rüdiger Weber, and Patrick Weiss. 2024. "Methodological Uncertainty in Portfolio Sorts." SSRN Working Paper. https://doi.org/10.2139/ssrn.4164117.

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

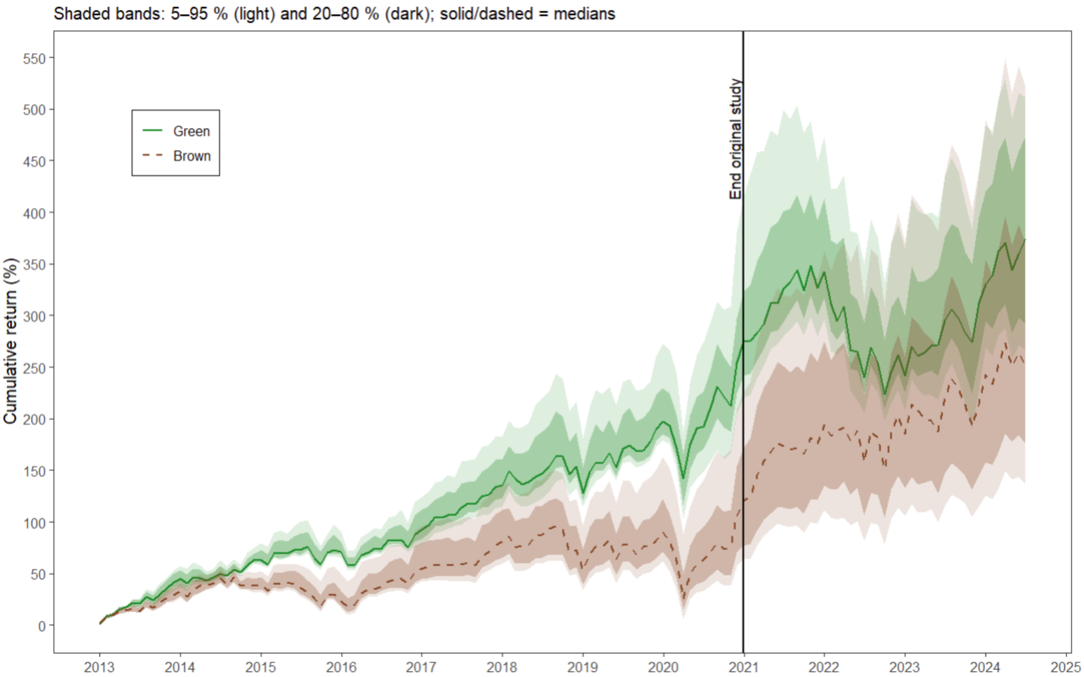
Xiao, Yuchao, Robert Faff, Philip Gharghori, and Darren Lee. 2013. "An Empirical Study of the World Price of Sustainability." *Journal of Business Ethics* 114 (2): 297–310. https://doi.org/10.1007/s10551-012-1342-2.

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

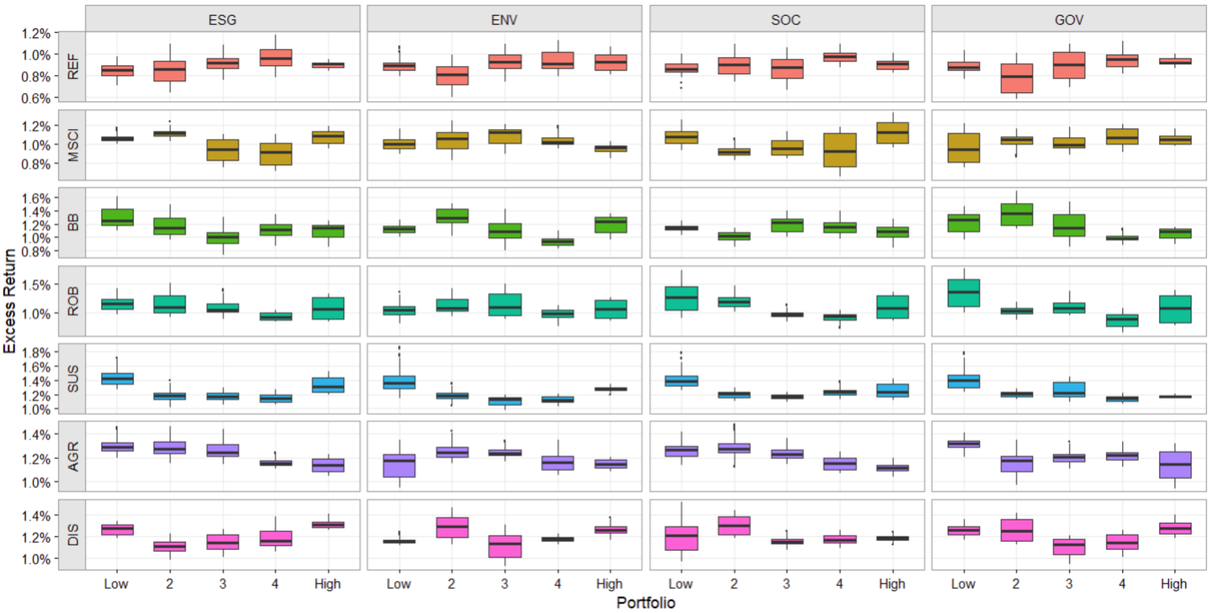


Explanation: Panel A of this Appendix features the value-weighted returns on green and brown portfolios created using the single-specification method of Pástor et al. (2022) over the extended period November 2012-June 2024. The cumulative return difference (CRD) between the two portfolios was 210%. Panel 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%.

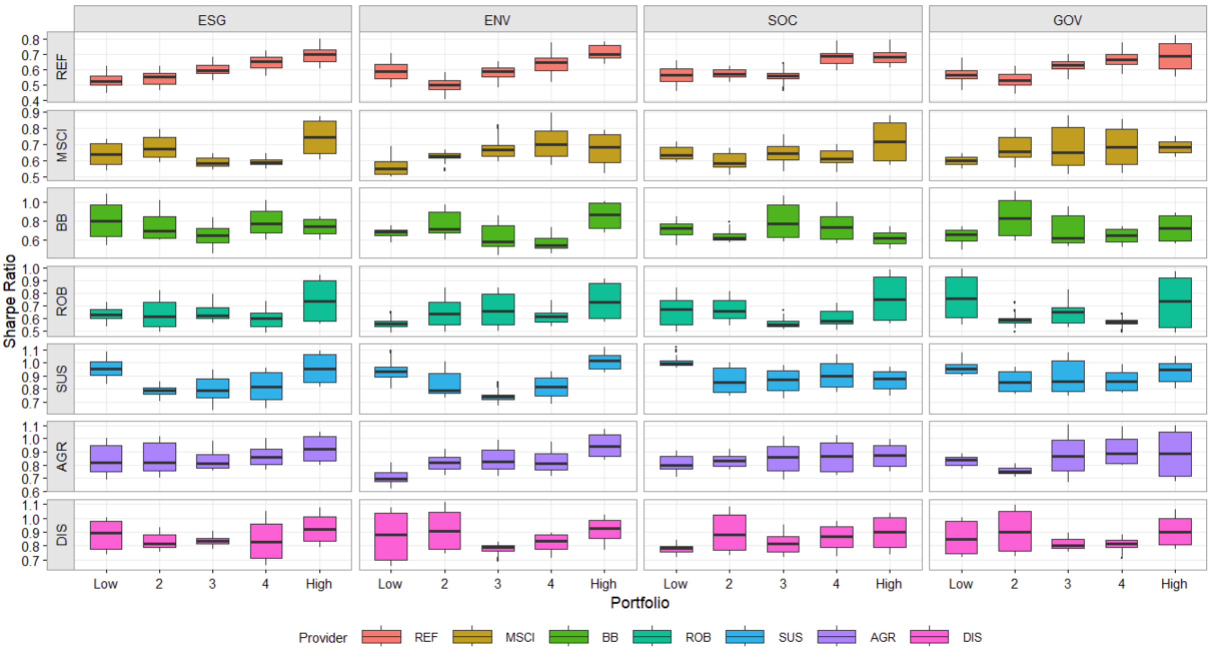
Interpretation: The single-specification replication shows clear green portfolio outperformance; however, our multi-specification estimates indicate that return difference between green and brown portfolios is sensitive to the chosen specification, especially after the original study period ends. Post-December 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 (%)



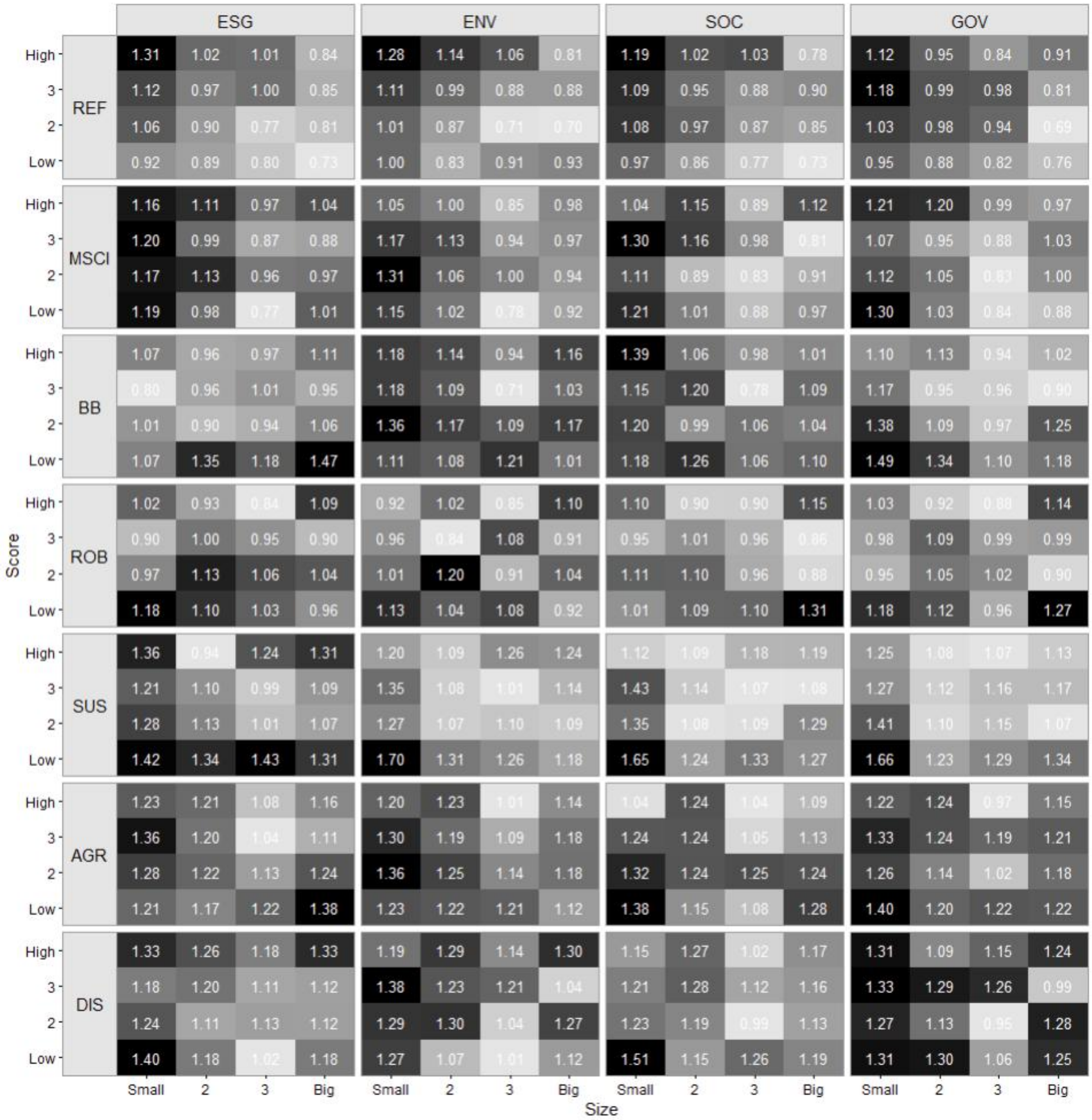
Panel B: Sharpe ratios



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.

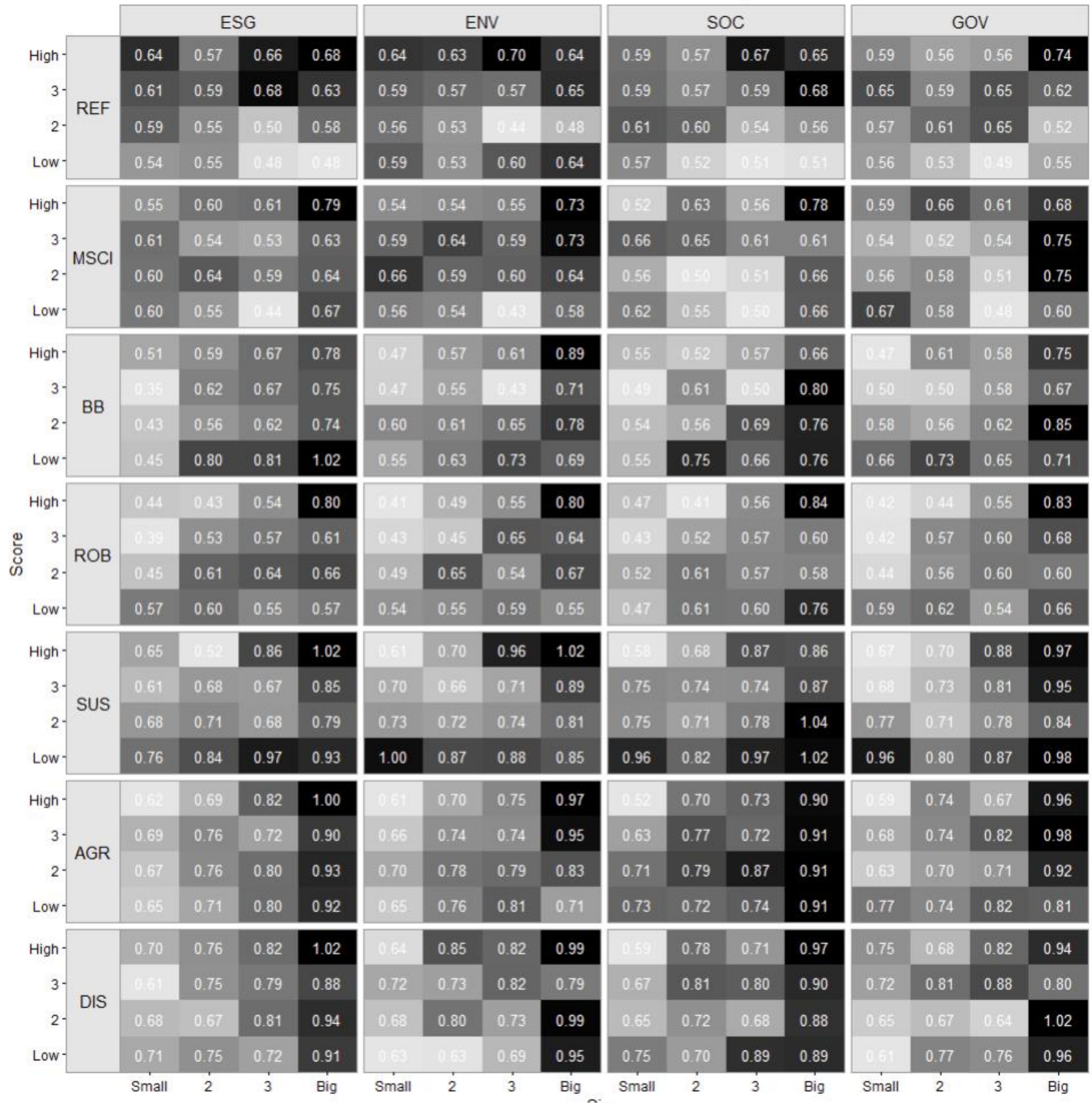
Appendix C. Double-sorted 4x4 portfolios: excess returns and Sharpe ratios

Panel A: Excess Returns (%)



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

Panel B: Sharpe ratios



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).

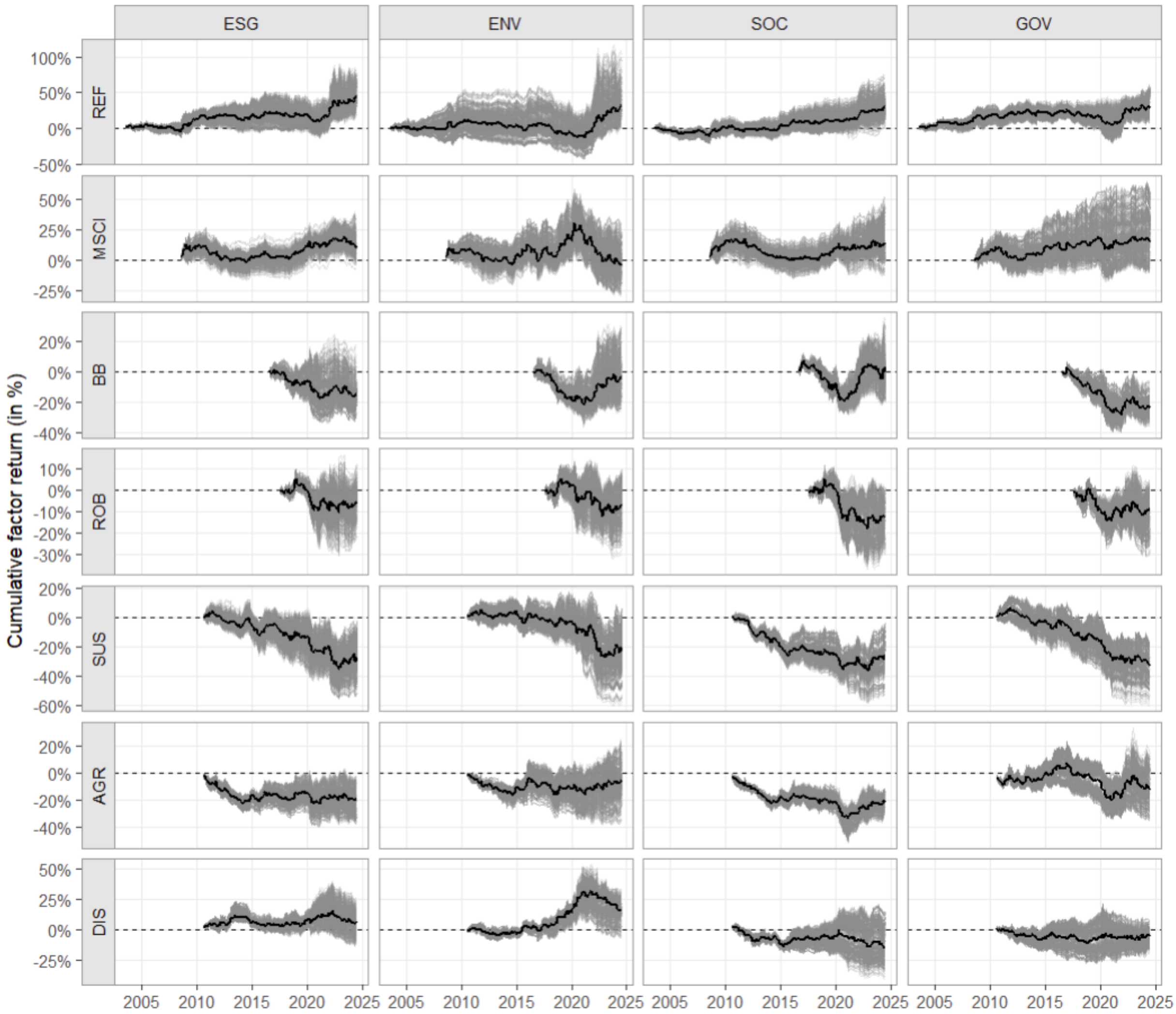
Interpretation: Median excess returns and Sharpe ratios of double-sorted 4×4 rating×size portfolios show no consistent monotonic increase as rating score improves. The relationships between rating scores and excess returns/Sharpe ratios differ across providers and rating types; though to some extent,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

[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
2017	2516	2218	298	2170	2139	31	2698	2658	40	2245	452	1793	630	616	14
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 (1)	Rater 2 (2)	Replication (3)	Original (4)	R – O (5)	Replication (6)	Original (7)	R – O (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 (1)	Mean (2)	Std Dev (3)	10% (4)	25% (5)	Median (6)	75% (7)	90% (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 (1)	Mean (2)	Std Dev (3)	10% (4)	25% (5)	Median (6)	75% (7)	90% (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 (1)	Mean (2)	Std Dev (3)	10% (4)	25% (5)	Median (6)	75% (7)	90% (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))

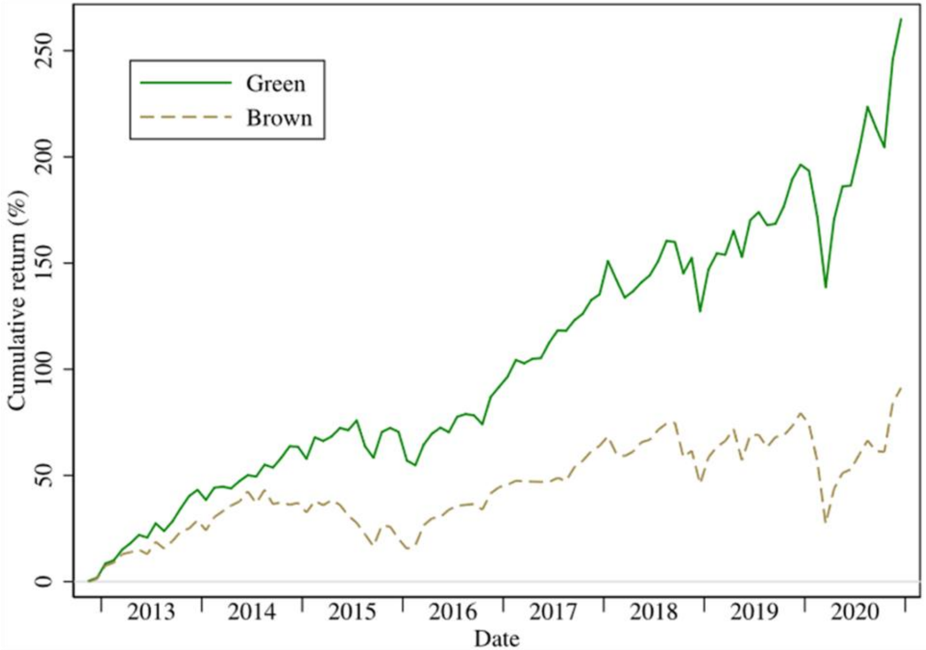
		Return (1)	CAPM (2)	FF3 (3)	C4 (4)	FF5 (5)
Original study	Return/Alpha	0.65	0.71	0.50	0.47	0.50
Nov 2012–Dec 2020	T-Value	**3.23**	**2.91**	**2.23**	**2.14**	**2.38**
Replication	Return/Alpha	0.66	0.73	0.52	0.50	0.51
Nov 2012–Dec 2020	T-Value	**3.76**	**3.27**	**3.14**	**3.04**	**3.25**
Replication	Return/Alpha	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 (1)	Original (2)	Rep: ESG (3)	Rep: ENV (4)	Rep: SOC (5)	Rep: GOV (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.

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.

[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 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	ESG		ENV		SOC		GOV	
	Original	Rep	Original	Rep	Original	Rep	Original	Rep
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(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.

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.

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 (1)	Portfolio (2)	Return (3)	N (4)	StD. (5)	SR (6)	CAPM (7)	FF3 (8)	C4 (9)	FF5 (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 (1)	Portfolio (2)	Return (3)	N (4)	StD. (5)	SR (6)	CAPM (7)	FF3 (8)	C4 (9)	FF5 (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 (1)	Portfolio (2)	Return (3)	N (4)	StD. (5)	SR (6)	CAPM (7)	FF3 (8)	C4 (9)	FF5 (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 (1)	Portfolio (2)	Return (3)	N (4)	StD. (5)	SR (6)	CAPM (7)	FF3 (8)	C4 (9)	FF5 (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.

Declaration of Own Contribution

The first research paper, “Financial Returns of Going Green: Evidence from MSCI Indices”, is co-authored with Thomas Brückner and Huong Dang. The authors contributed to this research project as follows. I was responsible for the software implementation, data preparation, formal analysis, and visualization. I jointly developed the conceptualization and methodology with Huong Dang. Thomas Brückner provided the MSCI index constituent data. I jointly wrote the paper with Huong Dang, while Thomas Brückner contributed the descriptions of the index variants.

The second research paper, “Investing in a Low-Carbon Transition: Carbon Footprint Savings with Green MSCI Indices”, is co-authored with Huong Dang and Thomas Brückner. The authors contributed to this research project as follows. I was responsible for the software implementation, data preparation, formal analysis, and visualization. I jointly developed the conceptualization and methodology with Huong Dang. Thomas Brückner provided the MSCI index constituent data. I jointly wrote the paper with Huong Dang, while Thomas Brückner contributed the descriptions of the index variants.

The third research paper “Small Deviations, Large Reductions: An Optimization Approach to Low-Tracking Error Decarbonized European Equity Portfolios” is a single-author paper.

The fourth research paper, “Sustainability Ratings, Equity Portfolio Performance, and Factor Models: Evidence from a Multi-Specification Approach”, is co-authored with Huong Dang and Manuel Brinkmann. The authors contributed to this research project as follows. The conceptualization was developed jointly by all three authors. I developed the methodology, including the multi-specification framework. The replication of existing studies was developed jointly by Huong Dang and me. I conducted the software implementation, data curation, formal analysis, and validation. Visualization was primarily my responsibility, with Manuel Brinkmann in a supportive role. I mainly wrote the paper together with Huong Dang. Manuel Brinkmann contributed to the writing of the introduction and supported the preparation of the Online Appendix. I jointly presented with Manuel Brinkmann this research project at the 2024 International Doctoral Seminar in Banking and Finance in Liechtenstein. The paper was also accepted for a poster session at the Frontiers of Factor Investing Conference 2026, where I presented it.

I am the first author of all four research papers.