In partnership for the goals? The level of agreement between SDG ratings

Tobias Bauckloh, Juris Dobrick, André Höck, Sebastian Utz, Marcus Wagner

Angaben zur Veröffentlichung / Publication details:

Bauckloh, Tobias, Juris Dobrick, André Höck, Sebastian Utz, and Marcus Wagner. 2024. "In partnership for the goals? The level of agreement between SDG ratings." Journal of Economic Behavior & Organization 217: 664–78. https://doi.org/10.1016/j.jebo.2023.11.014.

Journal of Economic Behavior and Organization 217 (2024) 664–678

Contents lists available at ScienceDirect

Journal of Economic Behavior and Organization

journal homepage: www.elsevier.com/locate/jebo

Research Paper

In partnership for the goals? The level of agreement between SDG ratings

Tobias Bauckloh [a], Juris Dobrick [b], André Höck [c], Sebastian Utz [d,e,g,*], Marcus Wagner [d,f,g]

[a] Centre for Financial Research Cologne, Faculty of Management, Economics and Social Sciences, University of Cologne, Germany
[b] Faculty of Economics and Management, University of Kassel, Germany
[c] EB – Sustainable Investment Management GmbH, Kassel, Germany
[d] Faculty of Business and Economics, University of Augsburg, Germany
[e] School of Finance, University of St. Gallen, Switzerland
[f] Bureau d'Economie Théorique et Appliquée, University of Strasbourg/CNRS, Strasbourg, France
[g] Centre for Climate Resilience, University of Augsburg, Germany

ARTICLE INFO

JEL code:
G1
C5
Q5

Keywords:
Sustainable development goals
Disagreement
Portfolio management
Ratings

ABSTRACT

This study analyzes the level of agreement of Sustainable Development Goal (SDG) ratings across five different rating providers. It documents a low level of agreement that is particularly pronounced for companies from the Energy, Healthcare, and Basic Materials sectors. Moreover, the low level of agreement is mostly driven by some individual SDGs. When analyzing implications, we find different return characteristics and risk factor exposures of portfolios sorted according to SDG ratings of different rating providers. Overall, our analyses show that current SDG ratings fail to provide a clear signal of companies' contribution to the SDGs which can have severe consequences for sustainability transitions and their financing.

1. Introduction

Companies are thought to have a high potential to contribute to achieving the 17 Sustainable Development Goals (SDGs) of the United Nations (van Zanten and van Tulder 2018). While this general potential has been acknowledged, companies have different response strategies to the demands outlined in the SDGs. These range from, on the one hand, potentially more business case-oriented approaches to more triple-bottom-line-oriented strategies on the other hand (Fiandrino et al. 2022; Scherer et al. 2013; van Zanten and van Tulder 2021). Next to differing response options, the literature has also pointed to a lack of standardized measures or measurement systems for a company's contribution to the SDGs or any of the 169 action targets into which the SDGs are further disaggregated (Berrone et al. 2023). These two factors combined make the assessment of companies' SDG contribution very challenging.

Partly in response to this challenge, financial service providers started to issue ratings that aim to more comprehensively and systematically measure the contribution of companies to the SDGs. SDG ratings specify whether the business activities of a company contribute positively or negatively to the SDGs. Through these ratings, companies can gauge their contribution to the SDGs and stakeholders, such as investors, can use them in relevant decisions. Therefore, these ratings might help allocate money to companies with a high SDG contribution, which could help to close the existing annual investment gap of about $4 trillion to achieve the goals by

* Corresponding author at: Department of Climate Finance, University of Augsburg, Universitaetsstrasse 16, Augsburg 86159, Germany.
E-mail address: sebastian.utz@wiwi.uni-augsburg.de (S. Utz).

https://doi.org/10.1016/j.jebo.2023.11.014
Received 28 April 2023; Received in revised form 10 November 2023; Accepted 13 November 2023
Available online 9 December 2023

2030 (United Nations 2022). However, a prerequisite is that these ratings give a clear signal as to which companies contribute to the SDGs, i.e., that the SDG ratings of different rating providers for the same company coincide. This study addresses this prerequisite and assesses the level of agreement between SDG ratings.

Anecdotal evidence suggests that SDG ratings of different rating providers vary substantially. For example, the pharmaceutical company Johnson & Johnson receives an aggregated SDG rating of 0.06 from MSCI and of 8.60 from ISS[1]. Both rating providers create company SDG ratings between −10 and 10, and the higher the SDG contribution is, the higher the rating. Moreover, both rating providers base their calculation, among other things, on the amount of a company's revenues that contribute to one or more SDGs. Thus, it is surprising that Johnson & Johnson is in the Bottom SDG group[2] of the MSCI company universe, but in the Top SDG group of the ISS company universe.

In our empirical analysis, we analyze the level of agreement between SDG ratings of five different rating providers. Descriptive statistics show that the three highest-ranked regions and sectors in the Top SDG groups differ substantially, although we study the same sample of companies for each provider. A test of how many companies are grouped into the same SDG group (Bottom, Middle, Top) by the providers shows that the average agreement lies mostly between 30 % to 50 %. To identify determinants of the disagreement (i.e., low level of agreement), we calculate two SDG rating disagreement measures and estimate cross-sectional regressions with a set of explanatory variables. The results document that the primary industry sector of a company explains a part of the variation in the disagreement measures. While the SDG ratings of the five rating providers show small disagreement in sectors such as Technology, Financials, Consumer Discretionary, and Telecommunications, the disagreement is particularly high in the Energy, Healthcare, and Basic Materials sectors. To illustrate, the difference between the highest and lowest SDG rating for companies from the Healthcare sector is about 1.42 times larger than for companies from the Telecommunications sector. Finally, we find that certain individual SDGs such as SDG 13 "Climate Action" drive the disagreement in the aggregated SDG rating. We conclude that SDG ratings currently do not provide a clear signal on a company's SDG contribution.

To assess further implications of the SDG rating disagreement for stakeholders, such as investors, we analyze investment outcomes of portfolios that are formed on the basis of SDG ratings. For every SDG rating provider, we calculate exposures of Bottom, Middle, and Top SDG group portfolios to common risk factors such as market, size, book-to-market equity, investment, and profitability. We find a large heterogeneity in risk exposures across the SDG group portfolios of individual SDG rating providers. The heterogeneity is especially pronounced for the size risk factor. Additionally, we calculate the returns and risk exposures of each rating provider's Top-minus-Bottom SDG group portfolio. These zero-investment portfolios differ substantially in terms of performance and exposures to the market, size, and investment risk factors. Therefore, investment outcomes of portfolios that are formed on the basis of SDG ratings depend on the chosen SDG rating provider.

This paper contributes to the literature on the characteristics of company sustainability measures (e.g., Berg et al. 2022; Dimson et al. 2020; Dorfleitner et al. 2015) by analyzing the disagreement between the recently introduced SDG ratings. It documents that SDG ratings of different providers disagree substantially. This is critical as these ratings might be used by companies to gauge their contribution to the SDGs and influence decisions of stakeholders. SDG ratings are a convenient possibility for stakeholders to align with SDGs and there is initial evidence that, unlike environmental, social, and governance (ESG) Ratings, SDG ratings capture their revealed sustainability preferences (van Zanten and Huij 2022). Agreement in SDG ratings might help to direct investment capital most efficiently towards solving the current challenges such as gender inequality (e.g., Brandts et al. 2021) as well as poverty alleviation and combating corruption (e.g., Han et al. 2022). However, our analyses reveal that SDG ratings, in their current shape, largely cannot fulfill this hope.

The remainder of this paper is structured as follows. Section 2 embeds our analysis in a theoretical framework. Section 3 describes our sample as well as SDG rating methodologies and explains the main variables of our analysis. Section 4 analyzes the level of agreement between the SDG ratings of five different providers. In Section 5, we identify determinants of the SDG rating disagreement in a cross-sectional regression model. Section 6 contains performance analyses for portfolios sorted with respect to SDG ratings. Finally, Section 7 concludes.

2. Theoretical framework

Stakeholders, such as investors, increasingly attach importance to companies operating sustainably and base their decisions on sustainability criteria (e.g., Krueger et al. 2020; Renneboog et al. 2008; van Duuren et al. 2016; Wen 2009; Wins and Zwergel 2015). If a company does not operate sustainably, investors might refrain from giving the company access to financial resources and customers might boycott a company's products and services. In this context, institutional theory suggests that companies respond to such pressures by integrating sustainability criteria in their business activities and communicate their contribution to desired sustainability goals such as the SDGs to support their legitimacy (e.g., DiMaggio and Powell 1983; Meyer and Rowan 1977). A company is considered legitimate if its activities do not violate the rules and values of its environment (Dowling and Pfeffer 1975; Suchman 1995). Deegan (2002) argues that legitimacy is critical to a company since it ensures access to important resources like a skilled workforce and decreases the probability of being targeted with retributions like fines or loss of sales. Thus, legitimacy is a crucial factor for a company's license to operate (Newson and Deegan 2002).

[1] Both ratings are from 2020. The methodologies of each rater are outlined in Section 3.

[2] We compute Bottom (Top) groups by taking the breakpoints of the lowest (highest) 30th percentile of the distributions of the SDG rating data of each provider.

Consequently, companies (stakeholders) need ways to communicate (assess) sustainability contributions. Possible ways are, for example, non-financial company disclosure (e.g., Deegan 2002; Reid and Toffel 2009; Reverte 2009) and companies' participation in sustainability initiatives (e.g., Bauckloh et al. 2023; Zerbini 2017). Moreover, commercial third-party sustainability rating agencies are an important intermediary between companies and stakeholders by providing company sustainability ratings such as ESG or SDG ratings. Given the popularity of such rating agencies among stakeholders (Berg et al. 2022), the question arises whether their ratings provide a correct and clear signal of a company's consideration of sustainability aspects? For example, several studies indicate that companies try to manage their sustainability rating by implementing ineffective actions that support a favorable one (e.g., Chatterji and Toffel 2010; Chelli and Gendron 2013; Clementino and Perkins 2021; Cornaggia and Cornaggia 2023; Searcy and Elkhawas 2012; Sharkey and Bromley 2015; Slager and Chapple 2016). Adding further complexity, ESG ratings from different sustainability rating agencies are based on different methodologies, leading to considerable disagreement (Berg et al. 2022; Chatterji et al. 2016; Christensen et al. 2022; Dimson et al. 2020; Dorfleitner et al. 2015). In this case, ESG ratings provide an unclear signal of a company's sustainability integration and impede appropriate decisions by stakeholders. Such disagreement between ESG ratings also has severe asset pricing implications (e.g., Avramov et al. 2022; Gibson Brandon et al. 2021; Serafeim and Yoon 2022).

Whilst the disagreement between ESG ratings can be related to different perceptions of what sustainability is, the SDGs translate the otherwise rather vague term "sustainability" into a well-defined framework (i.e., the 17 SDGs). Therefore, on the one hand, SDG ratings of different agencies could provide a clearer signal of a company's consideration of sustainability than ESG ratings and in doing so create legitimacy. On the other hand, however, Berg et al. (2022) point out that differences in ESG ratings are not only driven by differences in scope (understanding of sustainability), but also in measurement (specific indicators used for assessment by one rating agency). The latter can also lead to disagreement between SDG ratings. The same holds for ongoing discussions about which industries are considered unsustainable per se. For instance, Greenpeace planned to take the European Commission to court for labeling gas and nuclear energy as environmentally sustainable economic activities (Reuters 2023). Green- and bluewashing (Marquis et al. 2016) can also lead to disagreement between SDG ratings, if some rating methodologies detect green- and bluewashing better than others. Fig. 1 summarizes the theoretical framework.

Based on our framework, we test whether SDG ratings are a suitable means for companies (stakeholders) to ensure (assess) legitimacy by analyzing whether commercial ratings agencies create a clear signal of companies' SDG contributions.

3. Sample and data

3.1. Sample description

We obtain SDG ratings from five rating providers (MSCI, Inrate, Vigeo Eiris, ISS, and Robeco) for the year 2020. Table 1 shows sample statistics of the rating universe of each provider as well as the intersection sample ("All"), i.e., the sample of companies for which we have SDG ratings from all five rating providers. The original rating universes comprise 8,271 (MSCI), 1,986 (Inrate), 4,280 (Vigeo Eiris), 6,128 (ISS), and 7,998 (Robeco) companies with non-missing data. The intersection of these individual rating universes includes 1,057 companies. The intersection sample is our main sample in the analysis. Similar to Berg et al. (2022), we conduct cross-sectional analyses without time-series. This means that our results show point-in-time evidence and do not provide insights into whether the SDG ratings have converged over time.

Panel A of Table 1 presents the distribution of the sample companies regarding sectors. The rating universe of each rating provider and the intersection sample show a similar sector distribution. Consumer Discretionary, Financials, and Industrials are the top 3 sectors with the highest proportion of companies within each sample. Therefore, the intersection sample "All" appears to reflect the sector distribution of the original universes of the rating providers well. Panel B contains the distribution of the sample companies regarding regions. The largest three regions in terms of coverage are Asia, Europe, and North America. While Asian companies amount to more than 40 % of the companies in the rating universes of MSCI and Inrate, about 45 % of the companies in the rating universe of ISS are located in North America. In the intersection sample "All", 33 % of the companies are from North America, 35 % from Asia, and 24 % from Europe. Companies from other regions (Africa, Latin America & the Caribbean, and Oceania) amount to about 10 % of the companies in each sample except for Robeco. Therefore, also with respect to the distribution of the regions, the intersection sample

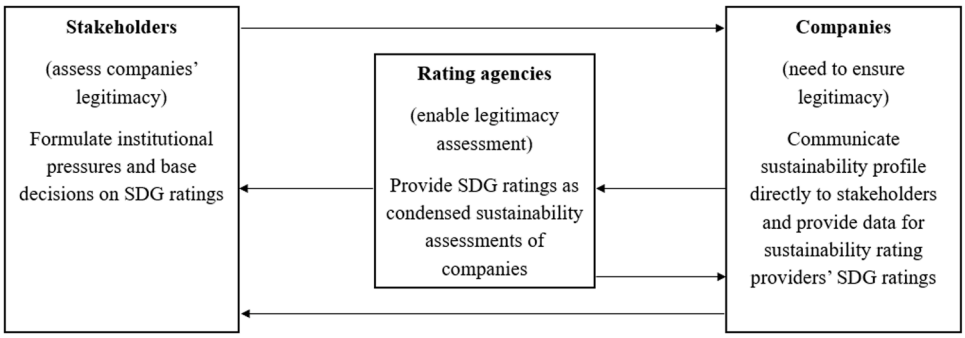

Fig. 1. Summary of theoretical framework.

Table 1
Sample statistics.

	MSCI	Inrate	Vigeo Eiris	ISS	Robeco	All
Number of companies	8,270	1,968	4,280	6,128	7,998	1,057
Panel A: Sectoral distribution						
Basic Materials	7.24	7.06	7.66	6.58	8.60	6.34
Consumer Discretionary	15.49	13.52	16.71	14.59	15.94	14.95
Consumer Staples	6.60	9.25	7.38	6.12	4.54	6.81
Energy	4.24	4.98	4.91	5.39	4.29	5.30
Financials	13.83	18.04	15.02	15.94	13.47	17.60
Healthcare	10.42	7.83	7.13	11.21	9.65	7.47
Industrials	17.95	15.29	18.46	17.27	19.14	17.60
Real Estate	8.02	4.78	6.24	7.20	7.69	4.64
Technology	9.50	9.65	8.18	8.40	10.35	9.84
Telecommunications	2.90	4.42	3.62	3.07	2.50	3.78
Utilities	3.80	5.18	4.70	4.24	3.83	5.68
Panel B: Regional distribution						
Africa	1.39	1.17	2.01	0.93	1.18	1.14
Asia	40.04	44.36	28.69	21.13	47.35	35.00
Europe	19.54	22.00	31.82	23.40	16.35	24.41
North America	30.90	23.63	24.74	45.77	30.02	33.40
Latin America & the Caribbean	4.09	6.86	3.79	3.04	2.21	3.41
Oceania	4.04	1.98	8.95	5.73	2.89	2.65
Panel C: Size quartile breakpoints (in $mn)						
25th percentile	966.53	4035.04	1562.92	1024.20	975.62	5697.05
50th percentile	2561.25	8236.27	4575.02	3305.43	2385.72	11499.26
75th percentile	7259.71	19415.82	12852.93	9256.76	6460.64	27494.33

This table shows sample statistics of the universes of the five different SDG rating providers (Columns "MSCI" "Inrate", "Vigeo Eiris", "ISS", and "Robeco"). Column 6 ("All") shows the statistics of the intersection of all five universes, i.e., the sample of companies that obtain an SDG rating from all five rating providers. The first row contains the absolute number of companies in each sample. Panel A (B) depicts the distribution of the companies with respect to sectors (regions). All values pertaining to sector and region are in percent. Panel C contains the samples' quartile breakpoints with respect to market capitalization in $mn.

"All" reflects the universes of each rating provider well.

Panel C of Table 1 shows the distribution of company size in each sample displayed in terms of quartile breakpoints of market capitalization in million USD ($mn), retrieved for all companies from Refinitiv Eikon as of 31st Dec 2020. The median company in the intersection sample has a market capitalization of 11,499 $mn. This number is substantially larger than the median size of the companies in the universes of the five rating providers. Thus, the intersection sample tends to consist of larger companies. This is reasonable since, similar to ESG ratings, most of the rating providers focus on assessing companies included in large stock market indices. If a difference in the rating universes exists, this difference could stem from smaller companies due, for instance, to regional biases of the rating provider. Although our samples differ in terms of company size, we include the largest companies and a large proportion of market capitalization in the "All" sample. For instance, although the MSCI sample includes almost seven times the number of companies, our intersectional "All" sample covers about 36.5 % of the market capitalization of the MSCI sample and 69 % of the Inrate sample.

3.2. SDG ratings

The five rating providers supply SDG assessments at different levels of aggregation for each company. In the main analysis of this paper, we study one aggregated SDG rating per company and rating provider. In the following, we explain the approaches of the five SDG rating providers related to our rating sample.

ISS provides SDG assessments for 15 different objectives on a scale from −10 to +10, where −10 (+10) indicates that 100 % of a company's operations as well as net sales are related to products/services that contribute negatively (positively) to one respective objective. The 15 objectives relate to social and environmental aspects of the SDGs. Out of these 15 single assessments, one overall rating is formed by taking the most extreme value(s) in either direction, i.e., positive or negative if there are only positive or negative ratings. If positive as well as negative ratings exist, the overall rating is computed by taking the sum of the highest and lowest values.

MSCI provides SDG assessments for each of the 17 SDGs on a scale from −10 to +10. The rating with respect to one SDG is calculated by averaging the SDG Product Alignment Rating and the SDG Operational Alignment Rating. A rating of −10 is assigned to a company that is strongly misaligned with an SDG. This can be the case when a company either generates over 50 % of its revenue from activities with adverse impact related to an SDG or if it is involved in major controversies related hereto. The SDG Product Alignment Rating measures the net contribution of a company's products and services. The SDG Operational Alignment Rating assesses the impact

of a company's operations.

Inrate maps a company's revenue to 300 standardized product and service segments. These product and service segments may contribute from "very negative" (−2) to "very positive" (+2) to each SDG, leading to an SDG net alignment with respect to each of the 17 SDGs in $mn.

Vigeo Eiris provides one overall SDG rating for each company in their universe. A company's "Behavior Score" encompasses several criteria pertaining to one of five behavioral themes that are relevant to a company's SDG assessment framework. These criteria are weighted and the company's performance on each of these is measured. Then, a decision tree determines the company's overall SDG contribution on a scale from "Highly Adverse" (−2) to "Highly Positive" (+2) by taking into account the company's behavioral performance versus its geographical peers, its involvement in controversial activities, and its involvement in sustainable goods and services. The involvement in sustainable goods and services is determined by several criteria pertaining to one of three product themes. Together with the five behavioral themes, there are overall eight themes that are relevant to the company's SDG assessment.

Robeco uses a three-step approach to compute a company's overall SDG rating by initially considering the impact of a company's products on broader society. This step is followed by an analysis of the company's operations and a final screening of potential controversies a company has been involved in. Thereafter, Robeco derives a rating for every single SDG on a scale from −3 to +3. An overall SDG rating is calculated by taking the most positive (negative) rating if a company has only positive (negative) ratings across all seventeen SDGs. If a company has positive and negative ratings across all SDGs, the overall SDG rating of a company is the most extreme negative rating.

Since MSCI and Inrate do not provide an aggregated SDG score, we follow the method from ISS to generate one aggregate SDG rating for them. We take the most extreme positive or negative rating in the seventeen SDGs if there are only positive or negative ratings for one company. If there are positive as well as negative ratings for one company, we take the sum of the most extreme positive and negative rating.[3]

Due to differences in scaling of the SDG ratings and to allow for a better comparison, we apply z-scoring and calculate a standardized SDG rating $SDG^z_{i,j}$ for each company i from each rating provider j

$$SDG^z_{i,j} = \frac{SDG_{i,j} - \mu_j}{\sigma_j} \tag{1}$$

where $SDG_{i,j}$ is the aggregated SDG rating of company i provided by rating provider j, μ_j is the cross-sectional mean of the aggregated SDG ratings of rating provider j and σ_j is the standard deviation of the aggregated SDG rating distribution of rating provider j. As a result, the distributions of the standardized SDG rating $SDG^z_{i,j}$ have a mean of 0 and a standard deviation of 1 for each rating provider and are therefore directly comparable.

Table 2 contains descriptive statistics of the standardized SDG rating $SDG^z_{i,j}$ and sample statistics for subsamples that we grouped with respect to SDG ratings for every rating provider. Rows starting with "Bottom" show the statistics of the subsample of companies with the 30 % lowest $SDG^z_{i,j}$ values (Bottom SDG group), rows starting with "Top" show the statistics of the subsample of companies with the 30 % highest $SDG^z_{i,j}$ values (Top SDG group). Correspondingly, the group in the middle represents the section in between (Middle SDG group). Moreover, Vigeo Eiris and Robeco have the specific case of a zero-standard deviation of the ratings in the Middle and Top SDG groups, respectively. For Vigeo Eiris this is due to the discreteness of the SDG ratings with only five different levels such that the Middle and the Top SDG groups only contain one single level of SDG ratings and thus no variability. Regarding Robeco, there is no variability in the SDG ratings in the Middle SDG group since there is only one rating level for this group. We address this issue by taking different aggregation methods (see online Appendix) as well as by taking different breakpoints to construct the groups. The main results for different approaches to group the companies remain qualitatively unaltered.

Table 2 additionally contains descriptive sample statistics such as the Top 3 countries, regions, and sectors. For instance, the column "Top 3 Countries" contains the list of countries (where the companies are headquartered) that occur with the highest frequency in the respective SDG group. The order of the country abbreviations indicates the order in the top 3. Thus, the US is the country with the highest number of companies in the Bottom SDG group of the MSCI SDG ratings. The second-most companies are from Canada, and the third-most companies are from Japan. Except for the Middle SDG group of the Vigeo Eiris Panel, US companies take the most prominent position across all groups and across all rating providers, followed by Japanese companies. This can be explained by the dominant share of North American (and here particularly US) companies in the overall sample. Concerning sectors, there seems to be more disagreement between the five rating providers: Energy and Industrials are the most represented sectors in the Bottom SDG group of MSCI and Vigeo Eiris, respectively, whereas it is Consumer Discretionary for ISS, Inrate, and Robeco. Companies in the Healthcare sector do well in all methodologies except MSCI. Healthcare is among the Top 3 sectors across all four remaining raters, although Healthcare only accounts for 7.5 % of the entire sample.

[3] For robustness, we compute aggregate SDG ratings in several ways: 1. We take the arithmetic mean/sum across all seventeen SDGs; 2. we use the method of Robeco. Furthermore, we apply the 1. and 2. approach to the raters that already provide aggregate SDG scores themselves, except for Vigeo Eiris. This is due to the lack of Vigeo Eiris ratings on a single SDG basis. Finally, we also apply the ISS aggregation method to all raters. We conduct all analyses for all five robustness sets. The central results of this study are the same across all sets of SDG aggregation. The results of the robustness tests are available in the online Appendix.

Table 2
Descriptive statistics and sample statistics per SDG rating group.

SDG group	N	Mean	SD	Min	Max	Top 3 countries	Top 3 regions	Top 3 sectors
					MSCI			
Bottom	340	−1.05	1.11	−3.17	−0.07	US, CA, JP	North America, Asia, Europe	Energy, Consumer Discretionary, Industrials
Middle	416	0.26	0.12	0.09	0.40	US, JP, CA	Asia, North America, Europe	Industrials, Financials, Consumer Discretionary
Top	301	0.82	0.32	0.56	2.11	US, JP, TW	Asia, Europe, North America	Financials, Industrials, Technology
					Inrate			
Bottom	317	−1.16	0.38	−1.72	−0.66	US, CA, JP	North America, Asia, Europe	Consumer Discretionary, Energy, Basic Materials
Middle	423	−0.02	0.28	−0.65	0.35	US, JP, KR	North America, Asia, Europe	Technology, Industrials, Financials
Top	317	1.18	0.59	0.35	1.98	US, JP, GB	Asia, North America, Europe	Financials, Healthcare, Industrials
					Vigeo Eiris			
Bottom	718	−0.54	0.71	−1.92	−0.01	US, JP, CA	North America, Asia, Europe	Industrials, Consumer Discretionary, Financials
Middle	271	0.94	0.00	0.94	0.94	JP, US, TW	Asia, Europe, North America	Financials, Technology, Industrials
Top	68	1.90	0.00	1.90	1.90	US, GB, FR	North America, Europe, Asia	Healthcare, Industrials, Consumer Discretionary
					ISS			
Bottom	321	−1.04	0.71	−2.62	−0.09	US, CA, JP	Asia, North America, Europe	Consumer Discretionary, Industrials, Consumer Staples
Middle	422	0.02	0.06	−0.06	0.17	US, JP, TW	Asia, North America, Europe	Financials, Industrials, Consumer Discretionary
Top	314	1.03	0.82	0.2	2.60	US, JP, GB	North America, Asia, Europe	Healthcare, Technology, Industrials
					Robeco			
Bottom	432	−1.00	0.73	−2.32	−0.29	US, JP, CA	Asia, North America, Europe	Consumer Discretionary, Industrials, Basic Materials
Middle	369	0.39	0.00	0.39	0.39	US, JP, TW	Asia, North America, Europe	Financials, Technology, Industrials
Top	256	1.13	0.19	1.07	1.75	US, JP, CH	North America, Asia, Europe	Healthcare, Financials, Industrials

This table shows descriptive statistics for SDG groups built with SDG ratings from 2020 in the "All" sample. N, Mean, SD, Min, and Max denote number of companies, arithmetic mean, standard deviation, minimum, and maximum values of the standardized SDG ratings, respectively. Top 3 countries, Top 3 regions, and Top 3 sectors denote the top 3 countries, the top 3 regions, and the top 3 sectors per group. The abbreviations CA, CH, FR, GB, JP, KR, TW, and US denote Canada, Switzerland, France, Great Britain, Japan, Korea, Taiwan, and the US, respectively.

4. Level of agreement between SDG ratings

To analyze the level of agreement between different rating providers on a company's SDG performance, we first present results for a general agreement indicator on a measurement construct provided by different actors, namely Krippendorff's alpha, in Table 3. Krippendorff (1998) suggests that a value higher than 0.8 indicates agreement between different rating providers on the construct being measured, with a minimum value of 0.667 recommended to at least be able to make rough statements. We find that all Krippendorff's alpha values are well below the recommended values, with a value of 0.43 for all five rating providers (Panel A). Combinations of 4, 3, and 2 (Panel B, C, and D) rating providers mostly yield values between 0.3 and 0.5, with the values for Inrate and ISS and ISS and Robeco standing out at 0.65 and 0.58, respectively.[4] Thus, in general, the level of agreement between different rating

[4] To get a better picture of how Krippendorff's alphas behave for ratings with generally high agreement, we calculate Krippendorff's alphas for Moody's and Fitch credit ratings for all bonds with non-missing data issued in 2020 for the companies in the "All" sample. For a total of 275 rated bonds, we obtain a Krippendorff's alpha value of 0.92.

Table 3
Krippendorff's alpha.

Panel A: All raters	
	0.43
Panel B: Four raters (the named rater is the rater omitted in the calculation)	
- MSCI	0.46
- Inrate	0.39
- Vigeo Eiris	0.50
- ISS	0.39
- Robeco	0.40
Panel C: Three raters (the named raters are the raters omitted in the calculation)	
- MSCI and Inrate	0.41
- MSCI and Vigeo Eiris	0.60
- MSCI and ISS	0.40
- MSCI and Robeco	0.44
- Inrate and Vigeo Eiris	0.45
- Inrate and ISS	0.34
- Inrate and Robeco	0.35
- Vigeo Eiris and ISS	0.46
- Vigeo Eiris and Robeco	0.35
- ISS and Robeco	0.35
Panel D: Pairwise (the named raters are considered)	
MSCI and Inrate	0.42
MSCI and Vigeo Eiris	0.31
MSCI and ISS	0.38
MSCI and Robeco	0.40
Inrate and Vigeo Eiris	0.32
Inrate and ISS	0.65
Inrate and Robeco	0.56
Vigeo Eiris and ISS	0.34
Vigeo Eiris and Robeco	0.31
ISS and Robeco	0.58

This table shows Krippendorff's alphas for standardized SDG ratings. Panel A shows Krippendorff's alpha for the sample of companies with SDG ratings available for all five raters, Panel B shows Krippendorff's alpha for all possible combinations of four out of the five rating providers, where for instance "– MSCI" indicates the case when MSCI is not considered. Panel C shows Krippendorff's alpha values for all possible combinations of three raters. Panel D shows all possible pairwise combinations between the rating providers.

providers on a company's SDG performance is low and comparable to the level of agreement between ESG ratings (see Berg et al. 2022).[5]

Next, we conduct a more in-depth analysis and look at SDG group company matches between different rating providers. Results are provided in Table 4. As in Table 2, we divide all companies included in our "All" sample into the Bottom, Middle, and Top SDG groups for each rating provider. We then calculate the share of companies in the SDG groups of one rating provider that are also included in the respective SDG groups of the other four rating providers. For example, in Panel A, 50.00 % (84.41 %; 50.88 %; 58.53 %) of the companies included in the Bottom SDG group of MSCI are included in the Bottom SDG group of Inrate (Vigeo Eiris; ISS; Robeco). 41 of 60 percentual matches are between 30 % and 60 %, with some outliers upwards as well as downwards. The highest average agreement is between Inrate and ISS, with an agreement between 50 % and 70 % in the Bottom and Top SDG groups respectively. Furthermore, some high values such as 85.29 % between Vigeo Eiris and ISS in the Top SDG group stand out. However, the fact that Vigeo Eiris only has 68 companies in the Top SDG group puts the value somewhat in a different perspective: considering that about 58 of the 68 companies are in ISS' Top SDG group, which corresponds to roughly 18 % of all companies in the Top SDG group of ISS, the high value of 85.29 % is in fact somewhat misleading at first glance. In Panel B, we illustrate the percentual match between all possible rating provider pairs across all groups. That is, 45.7 % of all companies were assigned to the same SDG group by MSCI and Inrate. For all possible rating provider pairs, we achieve a percentual match of around 45 %, except for all possible pairs of Inrate, ISS, and Robeco. This stresses our findings from Panel A. In general, Table 4 again indicates a low level of agreement between SDG ratings of different rating providers. There is only overall agreement on the SDG group across all raters for 196 companies of the "All" sample, which corresponds to 18.54 %. 52.04 % of these 196 companies belong to the Bottom SDG group, whereas for the Middle and Top SDG groups

[5] For robustness, we recalculate Table 3 and use a different z-scoring method, i.e., we apply z-scoring by both sector and region. For both approaches the resulting Krippendorff's alphas remain low, for the region approach almost unchanged and for the sector approach with all values decreasing between 0.1 and 0.2. The results are available upon request.

Table 4
Percentual matches.

Panel A: Per group			
	Bottom	Middle	Top
	MSCI		
N	340	416	301
Inrate	50.00	43.75	43.52
Vigeo Eiris	84.41	25.96	13.95
ISS	50.88	43.51	41.53
Robeco	58.53	34.86	36.21
	Inrate		
N	317	423	317
MSCI	53.63	43.03	41.32
Vigeo Eiris	86.12	30.02	15.46
ISS	65.62	53.66	56.47
Robeco	76.66	45.39	48.58
	Vigeo Eiris		
N	718	271	68
MSCI	39.97	39.85	61.76
Inrate	38.02	46.86	72.06
ISS	39.83	58.67	85.29
Robeco	48.33	49.45	64.71
	ISS		
N	321	422	314
MSCI	53.89	42.89	39.81
Inrate	64.80	53.79	57.01
Vigeo Eiris	89.10	37.68	18.47
Robeco	74.14	42.65	43.63
	Robeco		
N	432	369	256
MSCI	46.06	39.30	42.58
Inrate	56.25	52.03	60.16
Vigeo Eiris	80.32	36.31	17.19
ISS	55.09	48.78	53.52
Panel B: Entire match (N = 1,057)			
MSCI & Inrate	45.70		
MSCI & Vigeo Eiris	41.34		
MSCI & ISS	45.32		
MSCI & Robeco	42.86		
Inrate & Vigeo Eiris	42.48		
Inrate & ISS	58.09		
Inrate & Robeco	55.72		
Vigeo Eiris & ISS	47.59		
Vigeo Eiris & Robeco	49.67		
ISS & Robeco	52.51		

This table shows percentual matches between the rating providers for the "All" sample. Panel A shows the percentual share of companies that were assigned to the same SDG group (Bottom, Middle, Top) by pairs of rating providers. Panel B shows the percentual share of companies that were assigned to the same SDG group across all SDG groups and for all possible pairs of rating providers.

it is roughly 24 % each. Thus, although there is little overall consensus across all raters, the agreement is higher for companies with a low SDG assessment.[6]

5. Explaining the disagreement between SDG ratings

Next, similar to the approach of Dorfleitner et al. (2022) in the context of corporate social responsibility scandals, we identify determinants of the disagreement between the rating providers based on a broad set of regional, sectoral, company-level, and thematic variables. First, we explain our measures of disagreement in the overall SDG assessment by company size, sectors, and regions of

[6] Figure A.1 in the online Appendix provides a visual presentation of the heterogeneity of SDG ratings across rating providers.

company headquarter. Thereafter, we look at how the disagreement in individual SDGs drives the overall disagreement.

5.1. SDG disagreement measures and descriptive statistics

We construct two SDG disagreement measures, *sd* and *max–min*. For the first measure *sd*, we take all five standardized SDG ratings and calculate the standard deviation for each company in our matched "All" sample of 1,057 companies. The second measure *max–min* represents the difference of maximum and minimum standardized SDG rating for each company in our sample. We use both of these SDG disagreement measures as dependent variables in a cross-sectional regression model with a set of explanatory variables. As explanatory variables, we use logarithmized market value (log(MV)), a company's ICB sector, and the region of the company's headquarter from Refinitiv Eikon.[7] All monetary values are in US$.

Table 5 presents descriptive statistics for our two SDG disagreement measures (dependent variables) and firm size (continuous independent variable). The dependent variables *sd* and *max–min* have an arithmetic mean of 0.69 and 1.71, respectively, indicating that the mean standard deviation of the standardized SDG ratings of the five rating providers is on average 0.69 and the average difference between maximum and minimum values per company and across all five rating providers is 1.71. The skewness and kurtosis of the two variables indicate that both are slightly right-skewed, i.e., there are more lower values than higher values in both SDG disagreement measures, and the values are roughly normally distributed with only a few outliers.

5.2. Determinants of SDG disagreement measures

We estimate three sets of regression models. In the first set of models (see Columns (1) in Table 6), we show the estimated coefficients of a regression with the SDG disagreement measures as the dependent variables and log company size and ICB sector affiliation as independent variables. Standard errors (not reported) are clustered with respect to sector and region. The coefficients for log(MV) are significant at the 10 %-level in both regression specifications, indicating that the disagreement between rating providers rises as the company's size increases. A one standard deviation increase in the variable log(MV) corresponds to an increase of 0.013 (1.88 % compared to the sample average) in *sd* and 0.034 (1.99 % compared to the sample average) in the *max–min* in model. Moreover, out of ten ICB sectors, roughly half of the sectors have a significantly different level of agreement than our reference sector Telecommunications. Whereas the disagreement is significantly lower in the Technology sector (relative to the reference sector, Telecommunications), it is significantly higher in the following sectors: Basic Materials, Energy, Healthcare, and Utilities.

In the second set of models (Columns (2)), we add the company's headquarter region as another explanatory variable to our model. Explanatory power of variables included in the first model remains robust in most cases. However, region of headquarter does not seem to play an important role in explaining SDG rating disagreement, since none of the coefficients are significant.

In the third set of Columns (3), we include the disagreement measures of single SDG clusters (described in Table 7) as further independent variables and analyze how the disagreement in single SDGs drives the overall SDG disagreement between the raters. To do so, we conduct further regression analyses to explain the variables *sd* and *max–min* with the same measures of disagreement in a single SDG. Since not all providers offer ratings at each individual SDG level, we use the cluster assignment approach of ISS and form 15 SDG clusters related to social or environmental aspects for each rater to not reduce the number of available observations for our analysis. For Vigeo Eiris, we take the scores of the eight themes of the SDG Assessment framework and construct ratings for each single SDG by taking the arithmetic mean of all themes concerned with one SDG. The clusters are depicted in Table 7, with the first and second columns displaying cluster number and name, respectively. The third and fourth columns of Table 7 show the cluster designations by ISS and the matched SDGs from the remaining raters, respectively. SDGs 8, 9, and 17 are not assigned to any cluster. For each case in which more than one SDG is assigned to one cluster, we use the arithmetic mean of the affected SDGs to compute the cluster rating for each company. Then, we apply z-scoring to each cluster as outlined in Section 3. We obtain 15 standardized SDG cluster ratings for each of our 1,057 companies in our "All" sample and calculate *sd* and *max–min* for each cluster in line with our aggregate disagreement measures. Descriptives statistics are in Table A.1 in the appendix.

We use these cluster disagreement measures as independent variables along with the previously used company size proxy, sectors, and regions to investigate which cluster disagreements drive the aggregate disagreement controlling for company size, sector, and region. Columns (3) in Table 6 show the results of the regression analysis for our model both for *sd* and *max−min*. Adding the cluster disagreements to the model increases the adjusted R^2 to 30.12 and 29.56 for *sd* and *max−min*, respectively. This increase is highly significant compared with model (2) as indicated by our F-test.

We have two main takeaways from the regression results. First, the SDG ratings of companies from the sectors Healthcare, Basic Materials, and Energy show larger SDG disagreement measures. Second, beyond this structural element, SDG disagreement measures are additionally driven by the low level of agreement in some specific SDG clusters, which refer more to environmental (rather than social) aspects.

The sectoral effects suggest that on the one hand, more environmentally and socially damaging sectors, such as Basic Materials and Energy seem to be more difficult to assess. In such sectors incentives for companies to green- or bluewash are likely higher, than in relatively cleaner sectors. If raters are more or less skilled in picking up green- or bluewashing, than this should lead, *ceteris paribus*, to higher heterogeneity. An additional possible explanation are different perceptions of rating providers regarding the role of these

[7] Furthermore, we implement market to book equity, return on assets, leverage, and price to earnings as additional explanatory variables. However, we find these variables to have no explanatory power and we exclude them from our model for overview reasons.

Table 5
Descriptive statistics for the regression variables.

	Mean	SD	Min	Max	Skew	Kurt
sd	0.69	0.31	0.12	1.99	0.56	−0.06
max–min	1.71	0.76	0.28	4.32	0.46	−0.36
log(MV)	9.42	1.11	6.77	14.00	0.59	0.34

This table presents descriptive statistics of the variables used in the regression analysis in Table 6. The data is from the "All" sample. Mean, SD, Min, Max, Skew, and Kurt represent the arithmetic mean, standard deviation, minimum and maximum values, skewness, and kurtosis, respectively.

Table 6
Cross-sectional regression on the SDG disagreement.

	(1)		(2)		(3)	
	Dependent Variable		*Dependent Variable*		*Dependent Variable*	
	sd	*max–min*	*sd*	*max–min*	*sd*	*max–min*
Constant	0.519***	1.333***	0.584***	1.504***	0.403***	1.032***
log(MV)	0.013*	0.034*	0.015*	0.038*	0.004	0.010
Tech	−0.192***	−0.532***	−0.200***	−0.556***	−0.177***	−0.47***
BasM	0.161***	0.339**	0.148***	0.304**	0.113**	0.246**
ConD	0.012	0.012	0.010	−0.044	0.000	−0.034
ConS	0.107**	0.107	0.100**	0.158	0.075	0.148
Ener	0.277***	0.675***	0.265***	0.643***	0.119**	0.303**
Fin	−0.070	−0.232*	−0.079*	−0.255**	−0.045	−0.143
Hc	0.298***	0.619***	0.294***	0.607***	0.283***	0.642***
Ind	0.034	0.034	0.031	0.008	0.025	0.030
RealE	0.075	0.075	0.063	0.094	0.089	0.197
Util	0.262***	0.575***	0.254***	0.553***	0.131	0.277
Asia			−0.070	−0.176	−0.003	−0.016
Europe			−0.104	−0.255	−0.037	−0.094
Latin America & the Caribbean			−0.104	−0.270	−0.084	−0.225*
North America			−0.073	−0.173	0.001	−0.002
Oceania			0.080	0.216	0.113**	0.290**
Cluster 1					−0.017	−0.015
Cluster 2					0.009	0.006
Cluster 3					0.053**	0.029
Cluster 4					0.005	0.007
Cluster 5					0.002	0.000
Cluster 6					0.112***	0.123***
Cluster 7					0.026	0.032*
Cluster 8					0.001	0.003
Cluster 9					−0.027	−0.015
Cluster 10					−0.013	−0.012
Cluster 11					0.010	−0.001
Cluster 12					0.015	0.020
Cluster 13					0.131***	0.130***
Cluster 14					0.064***	0.068***
Cluster 15					−0.057**	−0.061**
Adj. R^2	20.51	19.86	21.14**	20.55**	30.12***	29.56***

This table presents the results for the cross-sectional regressions on the SDG disagreement between the five rating agencies of the "All" sample. The dependent variables are standard deviation (*sd*) and the range of maximum and minimum (*max–min*) of the standardized SDG ratings per company, respectively. Models (1), (2) and (3) display the results for regressions conducted with log(MV) plus sector (1), plus region (2), plus the 15 SDG clusters (3), respectively. Standard errors are clustered at region- and sector-level. The variable log(MV) denotes logarithmized market value. Tech, BasM, ConD, ConS, Ener, Fin, Hc, Ind, RealE, and Util denote Technology, Basic Materials, Consumer Discretionary, Consumer Staples, Energy, Financials, Healthcare, Industrials, Real Estate, and Utilities, respectively. The reference sector is Telecommunications. ***, **, * denote significance at the 0.01, 0.05, 0.1 level, respectively. Adj. R^2 is reported in percent.

sectors, for instance, in the transition to a low carbon economy. While some sustainability frameworks consider the Basic Materials and Energy sectors as important transitional sectors in which best-practice solutions are deemed sustainable, other frameworks define these sectors as per se unsustainable. For instance, while power generation using gas and nuclear is considered as being a transition technology in context of the EU Taxonomy, Greenpeace planned to take the European Commission to court over controversial gas and nuclear greenwashing (Reuters 2023).

If we look at Healthcare, the heightened heterogeneity could result from the fact that some rating approaches focus more on processes (in which case Healthcare would be linked to negative (especially environmental) SDG effects) whereas others emphasize

Table 7
SDG clusters.

No.	SDG cluster name	ISS variable	Inrate, MSCI, Vigeo Eiris, Robeco SDG variables
1	No poverty	"Alleviating poverty"	1 (No poverty)
2	Zero hunger	"Combating hunger and malnutrition"	2 (Zero hunger)
3	Ensuring health	"Ensuring health"	3 (Good health and well-being); 6 (Clean Water & Sanitation)
4	Quality education	"Delivering education"	4 (Quality education)
5	Gender equality	"Attaining gender equality"	5 (Gender equality)
6	Providing basic services	"Providing basic services"	1, 3, 4, 6, 7, 10 (Reduced inequalities), 11
7	Peace, justice and strong institutions	"Safeguarding peace"	16 (Peace, justice and strong institutions)
8	Achieving sustainable agriculture and forestry	"Achieving sustainable agriculture and forestry"	15 (Life on land), 2 (Zero Hunger)
9	Clean water and sanitation	"Conserving water"	6 (Clean water and sanitation)
10	Affordable and clean energy	"Contributing to sustainable energy use"	7 (Affordable and clean energy)
11	Sustainable cities and communities	"Promoting sustainable buildings"	11 (Sustainable cities and communities)
12	Responsible consumption and production	"Optimizing material use"	12 (Responsible consumption and production)
13	Climate action	"Mitigating climate change"	13 (Climate action)
14	Life below water	"Preserving marine ecosystems"	14 (Life below water)
15	Life on land	"Preserving terrestrial ecosystems"	15 (Life on land)
		Not assigned	8, 9, 17 (Decent work and economic growth; Industry, innovation and infrastructure; Partnership for the goals)

This table presents the mapping between the SDG clusters we use in the following analyses and the single SDGs. We follow the ISS SDG clusters. ISS provides us with SDG ratings for these clusters. For MSCI, Inrate, Vigeo Eiris, and Robeco, we calculate new SDG cluster ratings based on the SDG ratings provided by these rating agencies and the mapping of the single SDG values to the SDG clusters. In clusters, in which more than one SDG dimension is assigned to the cluster, i.e., clusters 3 ("Ensuring health"), 6 ("Providing basic services"), and 8 ("Providing basic services"), the SDG cluster ratings are calculated with the arithmetic mean of the single SDG ratings.

products (in which case Healthcare would appear to be contributing positively to many SDGs). This again would reduce the level of agreement.

Consistent with this explanation, the cluster disagreements that increase our dependent variables relate to clusters 6, 13, and 14 in Table 7. These clusters are strongly related to "Clean water and sanitation", "Affordable and clean energy", "Mitigating climate change" and "Life below water", which can also explain, why the industry effects for Basic Materials and Energy decline when including SDG clusters in the model, since these clusters pinpoint more narrow areas where green- or bluewashing is frequent. Furthermore, it can explain that the Utility sector which was significant in Columns (2) of Table 6, now becomes insignificant, since the environmental impact of this sector most strongly relates to energy and water issues, which are again picked up in a focussed manner by the above clusters. Thus, in relation to our theorizing, we confirm that company and rater characteristics jointly affect the level of agreement.

6. Implications of disagreement between SDG ratings for portfolio management

Our analysis so far shows a low level of agreement between rating providers concerning the SDG contribution of a company. To gain further insight in the implications of the disagreement for stakeholders, we study investment outcomes of portfolios that are built based on SDG ratings. More specifically, we conduct portfolio return regressions using the 5-factor model from Fama and French (2015). Daily developed market factor returns are obtained from Kenneth French's data library.[8] We estimate regressions for value-weighted portfolios of each SDG group (Bottom, Middle, Top) of each rating provider. To this end, we proceed as in the previous sections and use standardized SDG ratings to separate our matched sample of 1,057 companies into SDG groups for each rating provider. This process results in five sets of three portfolios with average standardized SDG ratings increasing from the Bottom SDG group portfolio to the Top SDG group portfolio.[9] In addition, we construct zero-investment (difference) portfolios by taking the difference of the Top and Bottom SDG group portfolios, i.e., of the portfolios with the highest and lowest standardized SDG ratings. We compute daily value-weighted portfolio returns for the year 2021 and run Fama/French 5-factor models. Some caution has to be applied when interpreting these results due to the short time series of one year and the influence of the Covid-19 pandemic on capital markets.

Table 8 presents the coefficients of the regression models. Significance tests are based on Newey-West robust standard errors. The intercept coefficients are in percent on a daily basis. Across all rating providers (except for MSCI), exposure to market risk declines from the Bottom SDG group portfolio to the Top SDG group portfolio, indicating that market risk is lower for high SDG companies.

[8] We thank Kenneth R. French for providing the data on https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.
[9] The average standardized SDG ratings per portfolio can be seen in Table 2.

Table 8
Regression results for value-weighted portfolios (daily).

Panel A: SDG group portfolios Portfolio	α	β_{MKT}	β_{SMB}	β_{HML}	β_{RMW}	β_{CMA}
			MSCI			
Bottom	0.01	0.83***	−0.26***	0.09*	0.09	−0.07
Middle	0.01	0.84***	−0.01	0.02	0.00	−0.13*
Top	0.03*	0.82***	−0.03	0.01	0.07	−0.24***
			Inrate			
Bottom	0.02	0.90***	0.21***	0.41***	−0.04	−0.13
Middle	0.02	0.91***	−0.27***	−0.15***	0.08	−0.29***
Top	0.01	0.67***	−0.17***	0.07	0.10*	0.09
			Vigeo Eiris			
Bottom	0.01	0.84***	−0.13***	0.15***	0.02	−0.20***
Middle	0.04**	0.88***	−0.10**	−0.05	0.20***	−0.13
Top	0.02	0.66***	−0.21***	−0.28***	−0.07	0.26*
			ISS			
Bottom	−0.01	0.84***	0.17***	0.37***	0.08	0.02
Middle	0.01	0.94**	−0.18***	0.13**	0.07	−0.47***
Top	0.04**	0.73***	−0.26***	−0.22***	0.04	0.09
			Robeco			
Bottom	0.00	0.89***	0.02	0.35***	0.07	−0.14*
Middle	0.02	0.81***	−0.26***	−0.18***	0.06	−0.17*
Top	0.04**	0.77***	−0.14***	0.00	0.06	−0.06
Panel B: Difference portfolios Portfolio	α	β_{MKT}	β_{SMB}	β_{HML}	β_{RMW}	β_{CMA}
Top−Bottom MSCI	0.02	−0.02	0.23***	−0.08	−0.02	−0.17
Top−Bottom Inrate	−0.01	−0.23***	−0.38***	−0.33***	0.14	0.23*
Top−Bottom Vigeo Eiris	0.01	−0.18***	−0.08	−0.43***	−0.09	0.46***
Top−Bottom ISS	0.04*	−0.11***	−0.42***	−0.59***	−0.03	0.07
Top−Bottom Robeco	0.04*	−0.12***	−0.16**	−0.34***	−0.01	0.08

This table shows the results for regressions with value-weighted SDG group portfolios built with SDG rating data from the year 2020 for the companies in the "All" sample (1,057 companies). The regressions are conducted with daily data for the year 2021. The portfolios are rebalanced at the end of June. All returns are continuously compounded. Panel A shows regression results for Bottom to Top SDG group portfolios. Panel B shows the regression results for the difference portfolios, where the Top SDG group portfolio is long and the Bottom SDG group portfolio is short (Top−Bottom). α, β_{MKT}, β_{SMB}, β_{HML}, β_{RMW}, β_{CMA} denote the intercept of the regression and factor exposures, respectively. The intercept coefficients are in percent on daily basis. ***, **, * denote significance at the 0.01, 0.05, and 0.1 level, respectively. For the three SDG group portfolios (Bottom, Middle, and Top), the null hypothesis for β_{MKT} is that β_{MKT} =1, all other coefficients are tested against 0.

With respect to the remaining risk factors, the results are substantially different. The Bottom SDG group portfolio is exposed to small caps for Inrate and ISS, whereas it is exposed to large caps for MSCI and Vigeo Eiris. The Top SDG group portfolio has a large cap exposure for all rating providers except MSCI. For the HML factor, all Bottom SDG group portfolios have a positive coefficient, suggesting that portfolios comprising the companies with low SDG ratings are mostly value stocks. For the Top SDG group portfolios, only Vigeo Eiris and ISS portfolios load significantly on HML, with a negative coefficient signaling an exposure toward growth stocks. Whereas there are almost no significant loadings on RMW, the loadings on CMA are negative for some cases.

All difference portfolios in Panel B of Table 8 have an alpha which is indistinguishable from zero at the 5 %-level. Therefore, there is no strong evidence for the existence of positive abnormal risk-adjusted returns with a strategy that was long in high SDG stocks and short in low SDG stocks for the investigated time series. The coefficients of the market risk factor of the long-short strategy portfolios are all significantly different from zero except for MSCI. This finding is in line with the results in Panel A, showing that there is a difference in market risk across SDG groups. The difference portfolios based on SDG ratings from Inrate, ISS, and Robeco show an exposure towards large companies with values of −0.38, −0.42, and −0.16, respectively, while the MSCI SDG rating portfolio has a positive and significant loading of 0.23, indicating an exposure to small caps. Consequently, depending on the rating provider, portfolios' systematic risks can be either towards large caps or towards small caps with fairly high exposures in both directions. E.g., the factor premium for SMB is −1.13 % for the year 2021 and thus the difference in returns between the MSCI and Inrate difference portfolios for the entire year was 0.69 % only due to different size exposures.[10] The remaining risk exposures are quite similar direction-wise but differ in magnitude and statistical significance. Consequently, if one applies an SDG rating integration approach to a

[10] To put this into perspective: in 2019 and 2020, these premiums were 3.40 % and −6.15 %, respectively, which would have led to more extreme return differentials between the portfolios in those years just because of the different factor exposures.

portfolio, depending on the rating provider, one is exposed to quite different risks and returns.

The results we obtain in the portfolio performance analysis for 2021 are reflected in the results of other studies on the link between sustainability ratings and financial performance. In (second-order) meta-analyses Friede et al. (2015) and Atz et al. (2023) find a majority of studies that indicate a higher performance of sustainable compared to conventional investments. Specifically in the context of SDG ratings, Martí-Ballester (2021) studies the financial performance of SDG-themed equity funds in China during the period from 2009 to 2019. While the study's main conclusion is that SDG-themed mutual funds achieve a similar financial performance as the market benchmarks, the results also show that healthcare mutual funds outperform energy, technology, and ethical mutual funds. Table 2 of our study illustrates that companies from the Healthcare sector are the largest proportion in the Top SDG group portfolios following the ISS and Robeco. These portfolios show significant positive abnormal returns (in terms of alpha, see Table 8) of the Top group portfolios (Panel A in Table 8) and zero-investment Top−Bottom portfolios (Panel B in Table 8).

7. Conclusion

Stakeholders, such as investors, increasingly consider companies' sustainability profiles in their decisions. Therefore, from a legitimacy perspective, companies have a great interest in communicating their sustainability profiles credibly to stakeholders. While the assessment of a company's sustainability profile is non-trivial, sustainability ratings such as SDG ratings might solve this problem and rating agencies therefore have a mediating role between a company and its stakeholders. A prerequisite for this is that SDG ratings of different providers furnish a clear signal of a company's SDG contribution. Against this background, this study analyzes the level of agreement between SDG ratings and finds that SDG ratings of different providers differ substantially. This result calls into question whether SDG ratings are a suitable instrument to align financial flows with the SDGs. Consequently, the strategy of many governments to fill funding gaps in the achievement of the SDGs with private funds is compromised, since SDG ratings are unlikely to provide a reliable basis for identifying SDG contributions.

Our study therefore provides starting points for further research in this area. First of all, since we analyze the level of agreement between SDG ratings only for one year and given that methodologies of sustainability ratings change over time, further research could evaluate whether the low level of agreement increases or decreases over time across longer sampling periods. Moreover, future research should further explore how to meaningfully aggregate different ratings, e.g., in terms of a "meta rating". Yet, in the absence of a reliable benchmark, it seems difficult to evaluate if averaging across all different ratings is a good approach. Equally, just leaving out the most deviant (positive or negative) rating (a strategy frequently pursued in government measurement procedures) may be misleading, since without a reliable benchmark, it cannot be assured that these ratings really are the most unreliable. A reliable aggregation approach that avoids misdirection by individual values probably requires a more differentiated analysis of the conditions of the actors involved, especially with regard to rating agencies and companies. Therefore, this could be another area for future research. Finally, from a financial perspective, it would be interesting to further evaluate to what degree the low level of agreement between SDG ratings has similar asset pricing implications as those of ESG ratings (Avramov et al. 2022).

Funding

The authors are pleased to acknowledge funding from the Stiftung Mercator for a project entitled 'Wissenschaftsplattform Sustainable Finance' (Rahmenprogramm Sustainable Finance, grant number 19026202).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jebo.2023.11.014.

Appendix

We present descriptive statistics of the disagreement measures for SDG ratings on SDG-cluster-level in Table A.1. The average disagreement in the clusters ranges between 0.6 and 0.8 in terms of *sd* and between 1.3 and 1.9 in terms of *max–min*[11] with the lowest

[11] The correlations for each cluster pair are > 0.99 in all cases. Correlations are available upon request.

disagreements in the clusters 4, 5, 7, 11, and 12. The disagreement in clusters 3, 6, 13, and 15 is somewhat higher. The standard deviation in the disagreement measures indicates that there is variation in the average disagreement across all clusters. For clusters 4, 5, 7, and 12, the maximum disagreement is quite high compared with other clusters, which, along with the high kurtosis values, indicates that the high average disagreement in these clusters might partly be driven by outliers. The skewness is positive for all cases, suggesting that there are more extreme large disagreements than small ones with respect to average disagreement.

Table A.1
Descriptive statistics of the disagreement measures for SDG ratings.

	sd						*max–min*					
	Mean	SD	Min	Max	Skew	Kurt	Mean	SD	Min	Max	Skew	Kurt
1	0.67	0.67	0.21	7.11	3.60	19.46	1.63	1.56	0.56	16.16	3.53	19.14
2	0.62	0.76	0.11	5.99	3.16	11.67	1.49	1.85	0.25	14.64	3.27	12.86
3	0.72	0.51	0.10	3.98	1.40	2.57	1.78	1.26	0.25	9.46	1.44	2.71
4	0.60	0.73	0.15	10.00	6.56	64.42	1.46	1.72	0.38	22.66	6.28	59.22
5	0.57	0.51	0.08	9.17	9.61	142.73	1.38	1.22	0.22	21.33	9.00	128.02
6	0.76	0.48	0.06	4.31	2.62	12.58	1.86	1.17	0.15	10.63	2.55	11.94
7	0.65	0.71	0.10	8.30	4.04	24.32	1.58	1.71	0.27	19.92	4.00	24.41
8	0.71	0.66	0.09	5.98	4.48	24.61	1.77	1.62	0.23	14.22	4.37	23.48
9	0.72	0.61	0.14	5.92	3.93	21.55	1.80	1.50	0.31	15.96	3.96	22.48
10	0.69	0.58	0.07	3.14	1.33	1.29	1.70	1.41	0.18	7.79	1.31	1.24
11	0.67	0.67	0.10	7.02	3.23	17.27	1.64	1.60	0.25	15.98	3.11	15.55
12	0.65	0.70	0.10	12.20	6.48	78.89	1.57	1.65	0.26	27.39	5.92	65.81
13	0.78	0.50	0.03	2.88	1.01	0.47	1.94	1.24	0.06	7.79	1.05	0.72
14	0.70	0.68	0.09	6.42	4.30	23.09	1.72	1.62	0.23	14.72	4.13	21.43
15	0.72	0.63	0.06	5.56	3.77	17.55	1.79	1.55	0.15	14.30	3.67	17.14

This table presents descriptive statistics of the variables used in the SDG cluster regression analysis for the companies in the "All" sample. The numbers in the first column refer to the 15 ISS SDG clusters depicted in Table 7. Descriptives statistics are presented for *sd* and *max–min*. Mean, SD, Min, Max, Skew, and Kurt represent arithmetic mean, standard deviation, minimum and maximum values, skewness and kurtosis, respectively.

References

Atz, Ulrich, van Holt, Tracy, Liu, Zongyuan Z., Bruno, Christopher C., 2023. Does sustainability generate better financial performance? review, meta-analysis, and propositions. J. Sustain. Finance Invest. 13 (1), 802–825. https://doi.org/10.1080/20430795.2022.2106934.

Avramov, Doron, Cheng, Si, Lioui, Abraham, Tarelli, Andrea, 2022. Sustainable investing with ESG rating uncertainty. J. Financ. Econ. 145 (2), 642–664. https://doi.org/10.1016/j.jfineco.2021.09.009.

Bauckloh, Tobias, Schaltegger, Stefan, Utz, Sebastian, Zeile, Sebastian, Zwergel, Bernhard, 2023. Active First Movers vs. Late Free-Riders? An Empirical Analysis of UN PRI Signatories' Commitment. J. Bus. Ethics 182 (3), 747–781. https://doi.org/10.1007/s10551-021-04992.

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

Berrone, Pascual, Rousseau, Horacio E., Ricart, J.E., Brito, Esther, Giuliodori, Andrea, 2023. How can research contribute to the implementation of sustainable development goals? An interpretive review of SDG literature in management. Int. J. Manag. Rev. 25 (2), 318–339. https://doi.org/10.1111/ijmr.12331.

Brandts, Jordi, El Baroudi, Sabrine, Huber, Stefanie J., Rott, Christina, 2021. Gender differences in private and public goal setting. J. Econ. Behav. Organ. 192, 222–247. https://doi.org/10.1016/j.jebo.2021.09.012.

Chatterji, Aaron K., Durand, Rodolphe, Levine, David I., Touboul, Samuel, 2016. Do ratings of firms converge? Implications for managers, investors and strategy researchers. Strateg. Manag. J. 37 (8), 1597–1614. https://doi.org/10.1002/smj.2407.

Chatterji, Aaron K., Toffel, Michael W., 2010. How firms respond to being rated. Strateg. Manag. J. 31 (9), 917–945. https://doi.org/10.1002/smj.840.

Chelli, Mohamed, Gendron, Yves, 2013. Sustainability ratings and the disciplinary power of the ideology of numbers. J. Bus. Ethics 112 (2), 187–203. https://doi.org/10.1007/s10551-012-1252-3.

Christensen, Dane M., Serafeim, George, Sikochi, Anywhere, 2022. Why is corporate virtue in the eye of the beholder? The case of ESG ratings. Account. Rev. 97 (1), 147–175. https://doi.org/10.2308/TAR-2019-0506.

Clementino, Ester, Perkins, Richard, 2021. How do companies respond to environmental, social and governance (ESG) ratings? Evidence from Italy. J. Bus. Ethics 171 (2), 379–397. https://doi.org/10.1007/s10551-020-04441-4.

Cornaggia, Jess, Cornaggia, Kimberly, 2023. ESG ratings management. SSRN J. https://doi.org/10.2139/ssrn.4520688.

Deegan, Craig, 2002. The legitimising effect of social and environmental disclosures - A theoretical foundation. AAAJ 15 (3), 282–311. https://doi.org/10.1108/09513570210435852.

DiMaggio, Paul J., Powell, Walter W., 1983. The iron cage revisited: institutional isomorphism and collective rationality in organizational fields. Am. Sociol. Rev. 48 (2), 147. https://doi.org/10.2307/2095101.

Dimson, Elroy, Marsh, Paul, Staunton, Mike, 2020. Divergent ESG ratings. J. Portfolio Manag. 47 (1), 75–87. https://doi.org/10.3905/jpm.2020.1.175.

Dorfleitner, Gregor, Halbritter, Gerhard, Nguyen, Mai, 2015. Measuring the level and risk of corporate responsibility—an empirical comparison of different ESG rating approaches. J. Asset Manag. 16 (7), 450–466. https://doi.org/10.1057/jam.2015.31.

Dorfleitner, Gregor, Kreuzer, Christian, Sparrer, Christian, 2022. To sin in secret is no sin at all: On the linkage of policy, society, culture, and firm characteristics with corporate scandals. J. Econ. Behav. Organ. 202, 762–784. https://doi.org/10.1016/j.jebo.2022.08.027.

Dowling, John, Pfeffer, Jeffrey, 1975. Organizational legitimacy: social values and organizational behavior. Pac. Sociol. Rev. 18 (1), 122–136. https://doi.org/10.2307/1388226.

Fama, Eugene F., French, Kenneth R., 2015. A five-factor asset pricing model. J. Financ. Econ. 116 (1), 1–22. https://doi.org/10.1016/j.jfineco.2014.10.010.

Fiandrino, Simona, Scarpa, Francesco, Torelli, Riccardo, 2022. Fostering social impact through corporate implementation of the SDGs: transformative mechanisms towards interconnectedness and inclusiveness. J. Bus. Ethics 180 (4), 959–973. https://doi.org/10.1007/s10551-022-05189-9.

Friede, Gunnar, Busch, Timo, Bassen, Alexander, 2015. ESG and financial performance: aggregated evidence from more than 2000 empirical studies. J. Sustain. Finance Invest. 5 (4), 210–233. https://doi.org/10.1080/20430795.2015.1118917.
Gibson Brandon, Rajna, Krueger, Philipp, Schmidt, Peter Steffen, 2021. ESG rating disagreement and stock returns. Financ. Anal. J. 77 (4), 104–127. https://doi.org/10.1080/0015198X.2021.1963186.
Han, Linsong, Li, Xun, Xu, Gang, 2022. Anti-corruption and poverty alleviation: evidence from China. J. Econ. Behav. Organ. 203, 150–172. https://doi.org/10.1016/j.jebo.2022.09.001.
Krippendorff, Klaus (1998): Content analysis. An introduction to its methodology. 14. [print.]. Newbury Park: Sage (The Sage CommText series, 5).
Krueger, Philipp, Sautner, Zacharias, Starks, Laura T, 2020. The importance of climate risks for institutional investors. Rev. Financ. Stud. 33 (3), 1067–1111. https://doi.org/10.1093/rfs/hhz137.
Marquis, Christopher, Toffel, Michael W., Zhou, Yanhua, 2016. Scrutiny, norms, and selective disclosure: a global study of greenwashing. Organ. Sci. 27 (2), 483–504. https://doi.org/10.1287/orsc.2015.1039.
Martí-Ballester, Carmen-Pilar, 2021. Analysing the financial performance of sustainable development goals-themed mutual funds in China. Sustain. Prod. Consum. 27, 858–872. https://doi.org/10.1016/j.spc.2021.02.011.
Meyer, John W., Rowan, Brian, 1977. Institutionalized organizations: formal structure as myth and ceremony. Am. J. Sociol. 83 (2), 340–363. https://doi.org/10.1086/226550.
Newson, Marc, Deegan, Craig, 2002. Global expectations and their association with corporate social disclosure practices in Australia, Singapore, and South Korea. Int. J. Account. 37 (2), 183–213. https://doi.org/10.1016/S0020-7063(02)00151-6.
Reid, Erin M., Toffel, Michael W., 2009. Responding to public and private politics: corporate disclosure of climate change strategies. Strateg. Manag. J. 30 (11), 1157–1178. https://doi.org/10.1002/smj.796.
Renneboog, Luc, Horst, Jenke ter, Zhang, Chendi, 2008. Socially responsible investments: institutional aspects, performance, and investor behavior. J. Bank. Finance 32 (9), 1723–1742. https://doi.org/10.1016/j.jbankfin.2007.12.039.
Reuters (2023): Greenpeace to sue EU over 'green label for gas and nuclear. Available online at https://www.reuters.com/business/sustainable-business/greenpeace-sue-eu-over-green-label-gas-nuclear-2023-02-09/, checked on 10/13/2023.
Reverte, Carmelo, 2009. Determinants of corporate social responsibility disclosure ratings by Spanish listed firms. J. Bus. Ethics 88 (2), 351–366. https://doi.org/10.1007/s10551-008-9968-9.
Scherer, Andreas Georg, Palazzo, Guido, Seidl, David, 2013. Managing legitimacy in complex and heterogeneous environments: sustainable development in a globalized world. J. Manag. Stud. 50 (2), 259–284. https://doi.org/10.1111/joms.12014.
Searcy, Cory, Elkhawas, Doaa, 2012. Corporate sustainability ratings: an investigation into how corporations use the Dow Jones Sustainability Index. J. Clean. Prod. 35, 79–92. https://doi.org/10.1016/j.jclepro.2012.05.022.
Serafeim, George, Yoon, Aaron, 2022. Stock price reactions to ESG news: the role of ESG ratings and disagreement. Rev. Account. Stud. 1–31. https://doi.org/10.1007/s11142-022-09675-3.
Sharkey, Amanda J., Bromley, Patricia, 2015. Can ratings have indirect effects? Evidence from the organizational response to peers' environmental ratings. Am. Sociol. Rev. 80 (1), 63–91. https://doi.org/10.1177/0003122414559043.
Slager, Rieneke, Chapple, Wendy, 2016. Carrot and stick? The role of financial market intermediaries in corporate social performance. Bus. Soc. 55 (3), 398–426. https://doi.org/10.1177/0007650315575291.
Suchman, Mark C., 1995. Managing legitimacy: strategic and institutional approaches. AMR 20 (3), 571–610. https://doi.org/10.5465/amr.1995.9508080331.
United Nations (2022): Promotion and protection of human rights: human rights situations and reports of special rapporteurs and representatives. Available online at https://digitallibrary.un.org/record/3988295?ln=en, checked on 3/20/2023.
van Duuren, Emiel, Plantinga, Auke, Scholtens, Bert, 2016. ESG integration and the investment management process: fundamental investing reinvented. J. Bus. Ethics 138 (3), 525–533. https://doi.org/10.1007/s10551-015-2610-8.
van Zanten, Jan Anton, Huij, Joop, 2022. Corporate sustainability performance: introducing an SDG score and testing its validity relative to ESG ratings. SSRN Journal. https://doi.org/10.2139/ssrn.4186680.
van Zanten, Jan Anton, van Tulder, Rob, 2018. Multinational enterprises and the sustainable development goals: an institutional approach to corporate engagement. J. Int. Bus. Policy 1 (3-4), 208–233. https://doi.org/10.1057/s42214-018-0008-x.
van Zanten, Jan Anton, van Tulder, Rob, 2021. Improving companies' impacts on sustainable development: a nexus approach to the SDGS. Bus. Strategy Environ. 30 (8), 3703–3720. https://doi.org/10.1002/bse.2835.
Wen, Shuangge, 2009. Institutional investor activism on socially responsible investment: effects and expectations. Bus. Ethics 18 (3), 308–333. https://doi.org/10.1111/j.1467-8608.2009.01565.x.
Wins, Anett, Zwergel, Bernhard, 2015. Private ethical fund investors across countries and time: a survey-based review. Qual. Res. Financ. Mark. 7 (4), 379–411. https://doi.org/10.1108/QRFM-10-2014-0030.
Zerbini, Fabrizio, 2017. CSR initiatives as market signals: a review and research agenda. J. Bus. Ethics 146 (1), 1–23. https://doi.org/10.1007/s10551-015-2922-8.