MIT Sloan School of Management

MIT Sloan School Working Paper 5822-19

Aggregate Confusion: The Divergence of ESG Ratings

Florian Berg, Julian F. Koelbel, and Roberto Rigobon

August 15, 2019

Aggregate Confusion: The Divergence of ESG Ratings.*

Florian Berg[1], Julian F. Koelbel[2,1], Roberto Rigobon[1]
[1]MIT Sloan
[2]University of Zurich

August 15, 2019

Abstract

This paper investigates the divergence of environmental, social, and governance (ESG) ratings. First, the paper documents the disagreement between the ESG ratings of five prominent rating agencies. The paper proceeds to trace the disagreement to the most granular level of ESG categories that is available and decomposes the overall divergence into three sources: Scope divergence related to the selection of different sets of categories, measurement divergence related to different assessment of ESG categories, and weight divergence related to the relative importance of categories in the computation of the aggregate ESG score. We find that measurement divergence explains more than 50 percent of the overall divergence. Scope and weight divergence together are slightly less important. In addition, we detect a rater effect, i.e., the rating agencies' assessment in individual categories seems to be influenced by their view of the analyzed company as a whole. The results allow investors, companies, and researchers to understand why ESG ratings differ.

*We thank Jason Jay, Kathryn Kaminski, Eric Orts, Robert Eccles, Yannick Le Pen, Andrew King, and Timo Busch for detailed comments on earlier versions of this paper. Also we thank the seminar participants at JOIM 2018 for their comments. Armaan Gori, Elizabeth Harkavy, Andrew Lu, and Francesca Macchiavello provided excellent research assistance. All remaining errors are ours. Correspondence to: Roberto Rigobon, MIT Sloan School of Management, MIT, 50 Memorial Drive, E62-520, Cambridge, MA 02142-1347, aggregateconfusion@mit.edu, tel: (617) 258 8374.

1 Introduction

Environmental, social, and governance rating providers[1] have become very influential institutions that inform a wide range of decisions in business and finance. Regarding business, 80 percent of CEOs believe that demonstrating a commitment to society is important[2] and look to sustainability ratings for guidance and benchmarking. An estimated USD 30 trillion of assets are invested relying in some way on ESG ratings[3]. There are also a large number of academic studies that rely on ESG ratings for their empirical analysis, arguing for example that good ESG ratings helped to prop up stock returns during the 2008 financial crisis (Lins et al., 2017).

However, ratings from different providers disagree dramatically (Chatterji et al., 2016). In our data set of five different ESG raters, the correlations between their ratings are on average 0.61, and range from 0.42 to 0.73. For comparison, credit ratings from Moody's and Standard & Poor's are correlated at 0.99[4]. This means that the information that decision-makers receive from rating agencies is relatively noisy. Three major consequences follow: First, ESG performance is unlikely to be properly reflected in corporate stock and bond prices, as investors face a challenge when trying to identify out-performers and laggards. Fama and French (2007) show that investor tastes can influence asset prices, but only when a large enough fraction of the market holds and implements a uniform non-financial preference. Therefore, even if a large fraction of investors have a preference for ESG performance, the divergence of the ratings disperses the effect of these preferences on asset prices. Second, the divergence frustrates the ambition of companies to improve their ESG performance, because they receive mixed signals from rating agencies about which actions are expected and will be valued by the market. Third, the divergence of ratings poses a challenge for empirical research as using one rater versus another may alter a study's results and conclusions. Taken together, the ambiguity around ESG ratings is an impediment to prudent decision-making that would contribute to an environmentally sustainable and socially just economy.

This paper investigates why sustainability ratings diverge. In the absence of a reliable measure of "true ESG performance," the next best thing is to understand what drives the differences of existing ESG ratings. In principle, there are two reasons why ratings diverge. They might diverge because rating agencies adopt different definitions of ESG performance, or they can differ because these agencies adopt different approaches to measuring ESG performance. Currently, it is unclear how much each of those two explain the observed dispersion in ratings. Our goal is to disentangle these sources of divergence by comparing ratings at the disaggregate level. To do so, we specify the ratings as consisting of three basic elements: (1) a scope of attributes, which denotes all the elements that together constitute the overall concept of ESG performance; (2) indicators that represent numerical measures of the attributes; and (3) an aggregation rule that combines the set of indicators into a single rating. Divergence between ratings can arise from each of these three elements, whereas differences regarding scope and aggregation rule represent different views about the definition of ESG performance, and differences regarding indicators represent disagreement about appropriate ways of measuring.

[1] ESG ratings are also called sustainability ratings or corporate social responsibility ratings. We use the terms ESG ratings and sustainability ratings interchangeably.

[2] https://www.accenture.com/hk-en/insight-un-global-compact-ceo-study

[3] GSIA 2018

[4] Since credit ratings are expressed on an ordinal scale, researchers usually do not report correlations. However, for the sake of illustration we used the data from Jewell and Livingston (1998), and calculated a Pearson correlation by replacing the categories with integers.

We identify three distinct sources of divergence. *Scope divergence* refers to the situation where different sets of attributes are used as a basis to form different ratings. For instance, attributes such as greenhouse gas emissions, employee turnover, human rights, and lobbying, etc., may be included in the scope of a rating. One rating agency may include lobbying, while another might not, leading to differences in the final aggregate rating. *Weight divergence* refers to the situation where rating agencies take different views on the relative importance of attributes and whether performance in one attribute compensates for another. For example, the human rights indicator may enter the final rating with greater weight than the lobbying indicator. Indeed, the scope and weight divergence could also be subsumed under *Aggregation divergence*, since excluding an attribute from a rating's scope is equivalent to including it with a weight of zero. Finally, *Measurement divergence* refers to the situation where rating agencies measure the same attribute using different indicators. For example, a firm's labor practices could be evaluated on the basis of workforce turnover, or by the number of labor cases against the firm. Both capture aspects of the attribute labor practices, but they are likely to lead to different assessments. Indicators can focus on processes, such as the existence of a code of conduct, or outcomes, such as the frequency of incidents. The data can come from various sources such as company reports, public data sources, surveys, or media reports, for example. We assume that the rating agencies are trying to measure the same attributes, but use different indicators. The final aggregate rating contains all three sources of divergence intertwined into one number. Our goal is to estimate to what extent to which each of the three sources drives the overall divergence.

Methodologically, we approach the problem in three steps. First, we categorize all indicators provided by different data providers into a common taxonomy of 64 categories. This categorization is a critical step in our methodology, as it allows us to observe the scope of categories covered by each rating as well as to contrast measurements by different raters within the same category. The taxonomy is an approximation, because most raters do not share their raw data, making a matching between identical indicators impossible. However, restricting the analysis to identical indicators would yield that the entire divergence is due to scope, i.e., that there is zero common ground between ESG raters, which does not reflect the real situation. Thus, we use a taxonomy that matches indicators by attribute. We created the taxonomy starting from the population of 641 indicators and establishing a category whenever at least two indicators from different rating agencies pertain to the same attribute. Indicators that do not pertain to a shared attribute remain uncategorized. As such, the taxonomy approximates the population of common attributes as granular as possible and across all raters. We calculate category scores for each rating by taking simple averages of the indicators that belong to the same category. Second, we estimate the original ratings to obtain comparable aggregation rules. Using the category scores established by the taxonomy, we estimate weights of each category in a simple non-negative linear regression[5]. The results are modeled versions of the real ratings which are comparable in terms of scope, measurement, and weight in the aggregation rule. Third, we calculate the contribution of divergence in scope, measurement, and weight to the overall ratings divergence using two different decomposition methods.

Our study yields three results. First, we show that it is possible to estimate the implied aggregation rule used by the rating agencies with an accuracy north of 90 percent on the basis of a common taxonomy. This demonstrates that although rating agencies take very different approaches, it is possible to approximate their aggregation rule with a simple linear weighted average. We also estimated the ratings using different methodologies, e.g. neural networks and random forests. The results are virtually identical. In the out-of-sample, the non-negative linear regression performed the best. Second, we find that 53 percent of the difference of the ratings stems from measurement

[5] Non-negative least squares constrain the coefficients to take either zero or positive values.

divergence, while scope divergence explains 44 percent, and weight divergence another 3 percent. In other words, 53 percent of the discrepancy comes from the fact that the rating agencies are measuring the same categories differently, and 47 percent of the discrepancy stems from aggregating common data using different rules. This means that for users of this data – financial institutions for instance – a sizable proportion of the discrepancy could be resolved by sharing the data on the indicator level and having a common procedure for aggregation. On the other hand, these results also suggest that different sustainability ratings cannot be made congruent simply by taking into account scope and weight differences. Therefore, standardizations of the measurement procedures are required. Third, we find that a significant portion of the measurement divergence is rater-specific and not category-specific, suggesting the presence of a *Rater Effect*[6]. In other words, a firm that performs well in one category for one rater, is more likely to perform well in all the other categories for that same rater. Inversely, if the same firm is evaluated poorly in one category by another rater, it is more likely to be evaluated poorly for all the other categories as well.

Our methodology relies on two main assumptions and we evaluate the robustness of each of them. First, the individual indicators are assigned to categories using our individual judgment. We needed to make several judgment calls to determine to which categories each individual indicator belongs to. To evaluate robustness, we sorted the indicators according to the Sustainability Accounting Standards Board taxonomy. The results are virtually identical. Second, the linear rule is not contingent on the industry or the sector where the firm operates. Many rating agencies openly state that their aggregation rules are different for different industries. In other words, they state that each industry has its own set of key issues. However, we impose the exact same aggregation procedure on all firms and all sectors. We need to implement these two approximations to be able to compare procedures from different rating agencies. These assumptions, however, seems to be relatively innocuous in our empirical strategy. We are able to get surprisingly good approximations of the final ratings in our procedures based on our taxonomy with simple linear rules[7].

Our paper extends a stream of research that has documented the divergence of ESG ratings (Chatterji et al., 2016, 2009; Semenova and Hassel, 2015; Dorfleitner et al., 2015; Delmas and Blass, 2010). Its key contribution is to explore the disaggregate data behind ESG ratings and explaining in detail the sources of divergence. Our study is related to research on credit rating agencies, in particular, those dealing with the question why credit ratings differ (Bongaerts et al., 2012; Güntay and Hackbarth, 2010; Jewell and Livingston, 1998; Cantor and Packer, 1997). Similar to Griffin and Tang (2011), we estimate the underlying rating methodologies to understand the differences in ratings. Additionally, our study is related to literature that is concerned with changing investor expectations, namely the integration of ESG performance in investment portfolios. Several studies show that there is a real and growing expectation from investors that companies perform well in terms of ESG

[6] The rater effect or rater bias has been extensively studied in sociology, management, and psychology, especially in performance evaluation. Shrout and Fleiss (1979) evaluate different correlation measures to assess the rater effects. This is one of the most cited papers in psychology in the area of the rater effect. In performance evaluation see Mount et al. (1997). They study how different ethnicity and positions within the organization peers, subordinates, and bosses rate each other, and how the ratings are affected by these categories. These are two of the most influential papers in this area. In finance and economics there are many papers that study the biases in credit rating agencies. See Griffin and Tang (2011) and Griffin et al. (2013) for papers studying the rater bias. See Fong et al. (2014) where the authors study how changes in the competition of analysts impacts the biases of credit rating agencies. They find that less competition tends to produce an optimistic bias of the rating agencies. In sum, both in psychology and in finance, one can find a long history of ratings exhibiting biases. Many of those biases are rating agency wide. Finally, Didier et al. (2012) discuss the rater effect within the mutual fund industry with a focus on international diversification.

[7] These errors are very small relative to the discrepancy observed. We explain more than 90 percent of the observed variation, while the discrepancy is an order of magnitude larger.

performance (Amel-Zadeh and Serafeim, 2018; Gibson and Krueger, 2018), especially with regard to risks associated with climate change (Krueger et al., 2018). ESG ratings are the operationalization of investor expectations regarding ESG, thus understanding ESG ratings improves the understanding of these changing investor expectations.

The paper is organized as follows: Section 2 describes the data sources, section 3 documents the divergence in the sustainability ratings from different rating agencies. Section 4 explains the way in which we structure the data and describes the data at the disaggregate level, in section 5 we decompose the overall divergence into the contributions of *Scope*, *Measurement*, and *Weight*. In that section we also document the rater effect. Finally, we conclude in section 6.

2 Data

ESG ratings first emerged in the 1980s as a service for investors to screen companies not purely on financial characteristics, but also on characteristics relating to social and environmental performance. The earliest ESG rating agency Vigeo-Eiris was established in 1983 in France and five years later Kinder, Lydenberg & Domini (KLD) was established in the US (Eccles and Stroehle, 2018). While initially catering to a highly-specialized investor clientele, such as faith-based organizations, the market for ESG ratings has widened dramatically, especially in the past decade. Estimates are that 30 trillion USD are invested in ways that rely on some form of ESG information (GSIA, 2018), a figure that has grown by 34 percent since 2016. As interest in sustainable investing grew, many early providers were acquired by established financial data providers, e.g. MSCI bought KLD in 2010, Morningstar bought Sustainalytics in 2010, ISS bought Oekom in 2018 (Eccles and Stroehle, 2018), and Moody's bought Vigeo-Eiris in 2019.

ESG rating agencies offer investors a way to screen companies for ESG performance in a similar way credit ratings allow investors to screen companies for creditworthiness. Yet, there are two important differences. First, while creditworthiness is relatively clearly defined as the probability of default, ESG performance is a concept that is still evolving. Thus, an important part of the service that ESG rating agencies offer is an interpretation of what ESG performance means. Second, while financial reporting standards have matured and converged over the past century, ESG reporting is in its infancy. While most major companies provide some form of ESG reporting, there are competing reporting standards and almost none of the reporting is mandatory. Thus, ESG ratings provide a service to investors by collecting and aggregating information across a spectrum of sources and reporting standards. As a result, ESG ratings agencies have considerable discretion in how to produce ESG ratings and may give different ratings to the same company.

We use the data of five different ESG rating providers: KLD[8], Sustainalytics, Vigeo-Eiris, Asset4, and RobecoSAM[9]. We approached each provider and requested access to not only the ratings, but also the underlying indicators, as well as documentation about the aggregation rules and measurement

[8] KLD, formerly known as Kinder, Lydenberg, Domini & Co., was acquired by RiskMetrics in 2009. MSCI bought RiskMetrics in 2010. The dataset was subsequently renamed to MSCI Stats as a legacy database. We keep the original name of the dataset to distinguish it from the MSCI dataset.

[9] Other data providers have been approached and our goal is to continue evaluating the sources of discrepancy among the most prominent rating agencies. RepRisk and MSCI provided us with the data, which we are still processing. We also requested the data from Oekom/ISS and TrueValueLabs. However, at the moment of writing this paper, we have not been granted access to their data.

protocols of the indicators. Together, these providers represent most of the major players in the ESG rating space as reviewed in Eccles and Stroehle (2018). We requested that the data set be as granular as possible.

The KLD dataset was the only one that did not contain an aggregate rating, even though it is frequently used in academic studies in aggregate form. The KLD data set provided only binary indicators for either "strengths" or "weaknesses" in seven dimensions. We created an aggregate rating for KLD by following the procedure that is chosen in most academic studies, namely summing all strengths and subtracting all weaknesses[10].

Table 1 provides some basic descriptive statistics of the data sets obtained from the different rating providers. The number of firms covered in 2014[11], the baseline year for our analysis, ranges from 1671 to 4566. The balanced sample showed in Table 1 contains 823 firms. The mean and ESG scores are higher in the balanced sample for all providers, indicating that the balanced sample tends to drop lower performing companies.

Table 1. Descriptive Statistics

Descriptive Statistics of full sample in 2014.

	Sustainalytics	RobecoSAM	Vigeo-Eiris	KLD	Asset4
Observations	4551	1668	2319	4295	4025
Mean	56.38	47.17	32.19	1.11	50.87
Standard Deviation	9.44	21.05	11.78	1.72	30.95
Minimum	29	13	5	-6	2.78
Median	55	40	31	1	53.13
Maximum	89	94	67	9	97.11

Descriptive Statistics of common sample in 2014.

	Sustainalytics	RobecoSAM	Vigeo-Eiris	KLD	Asset4
Observations	823	823	823	823	823
Mean	61.36	49.61	33.91	2.44	72.12
Standard Deviation	9.52	20.91	11.46	2.28	24.12
Minimum	36	13	6	-4	3.26
Median	61	46	33	2	80.47
Maximum	89	94	67	9	97.11

The descriptive statistics of the aggregate rating (ESG) in 2014 using the unbalanced and common sample for the five rating agencies KLD, Sustainalytics, Vigeo-Eiris, RobecoSAM, and Asset4.

Throughout the paper, we refer to three versions of this data set. The first two are the full and the common sample as shown in Table 1. The third version is the normalized common sample, where all variables are normalized to have zero mean and unit variance.

3 Measurement of Divergence

To motivate our analysis, we illustrate the extent of divergence between the different rating agencies. The first step is to compute the correlations of the ratings between different rating agencies at different levels of aggregation. In particular, on the ESG level as well as for the environmental, social, and governance dimensions. Second, we evaluate heterogeneity at the firm level. Simple correlations, although easy to understand, can mask important heterogeneity in the data. It is possible that low correlations are due to large disagreements in a small subset of the firms. To explore this possibility,

[10] See e.g. Lins et al. (2017)

[11] Although, we have data for other years, most of our analysis is cross sectional and therefore we concentrate on the year in which the greatest common sample.

we compute the average absolute distance to the median rating for each firm. Third, we explore the rankings of the firms. We determine the proportion of firms belonging to the top quantile, and the proportion that belongs to the bottom quantile. We then proceed with a thorough analysis for different quantiles. We develop a simple statistic called the Quantile Ranking Count. The conclusion of these four approaches is the same. There is a high level of disagreement across rating agencies, and the disagreement is quite heterogeneous.

3.1 Correlations of Aggregate Ratings

In this section we describe the correlations between the ESG ratings from different rating agencies. Table 2 shows the Pearson correlations between the aggregate ESG ratings, as well as the ratings in the separate environmental, social, and governance dimensions. Correlations of the ESG ratings are on average 0.61, and range from 0.42 to 0.73. The correlations of the environmental ratings are slightly higher than the overall correlations with an average of 0.65. The social and governance ratings have the lowest correlations with an average of 0.49 and 0.38, respectively. These results are consistent with Semenova and Hassel (2015), Chatterji et al. (2016), Dorfleitner et al. (2015), and Bouten et al. (2017).

KLD clearly exhibits the lowest correlations with all other raters, both for the ESG rating and for the individual dimensions. RobecoSAM and Vigeo-Eiris have the highest level of agreement between each other, with a correlation of 0.73.

Table 2. Correlation at aggregate ESG level and at E, S, and G level.

	SA - VI	SA - KL	SA - RS	SA- A4	VI - KL	VI - RS	VI - A4	KL - RS	KL - A4	RS - A4
ESG	0.73	0.53	0.68	0.67	0.48	0.71	0.71	0.49	0.42	0.64
E	0.70	0.61	0.66	0.65	0.55	0.74	0.66	0.58	0.55	0.70
S	0.61	0.28	0.55	0.58	0.33	0.70	0.68	0.24	0.24	0.66
G	0.55	0.08	0.53	0.51	0.04	0.78	0.77	0.24	-0.01	0.81
Econ	-	-	-	-	-	-	-	-	-	0.43

Correlations between the ratings on the aggregate level (E, S, G, and ESG) from the five different rating agencies are calculated using the common sample. The results are similar using pairwise common samples based on the full sample. SA, RS, VI, A4 and KL are short for Sustainalytics, RobecoSAM, Vigeo-Eiris, Asset4, and KLD, respectively.

The disagreement between ESG ratings is far larger than between credit ratings. Credit rating agencies use different data sources and procedures to evaluate the ability to pay as well as the willingness to pay of firms, governments, and individuals. These procedures and the data sources are not free of judgment. Nevertheless, we find a correlation of 98.6 percent between credit ratings from Moody's and Standard & Poor's. Since credit ratings are expressed on an ordinal scale, researchers usually do not report correlations. However, for the sake of illustration we used the data from Jewell and Livingston (1998), and calculated a Pearson correlation by replacing the categories with integers. The degree of disagreement between ESG ratings from different provider is thus far more pronounced. While credit rating agencies occasionally differ in their assessment one category upwards or downwards, ESG ratings disagree significantly more.

3.2 Heterogeneity in the Disagreement

The problem of correlations is that they are comparisons at the rating agency level. Correlations tend to obscure firm level differences. For example, two rating agencies can be weakly correlated because there is disagreement for every firm in the sample, or because there is agreement in a large set of firms and extremely large disagreement in a small set of firms. To evaluate this possibility we use the normalized common sample and compute the average absolute distance to the median rating for each firm. The normalized data indicates where the firm is located in the distribution of a particular rating agency. Even if the nominal ratings might differ, the placements in the distribution might be similar. This provides a firm-specific measure of disagreement[12]. To present the data we concentrate on the extremes of the distribution of the median average distance — the 100 firms with the highest agreement, and the 100 firms with the highest disagreement.

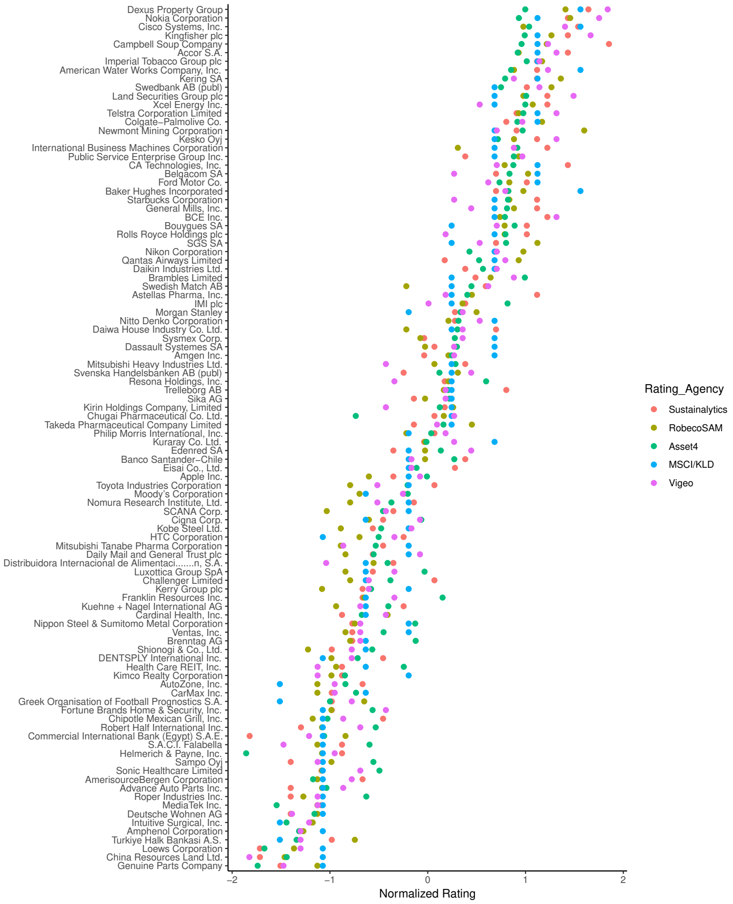


Figure 1. Comparison of firms' normalized scores for different rating agencies.
100 firms with the lowest median average distance within the normalized common sample (n=823). Firms are sorted by their median rating. Each rating agency is plotted in a different color. The vertical strings of blue dots are due to the fact that the KLD rating has only 14 unique values.

In Figure 1 we present a subset containing the 100 firms with the lowest average distance to the median, i.e., where the agreement between raters is greatest. To simplify the visualization, we rank the firms by their median, placing the best rated firm at the top and the worst rated firm at the bottom. The y-axis displays the firm's name, and the x-axis the normalized rating, reflecting how positively or negatively firms are rated among all five rating agencies. Each rating agency is depicted with a different colour[13].

The figure shows that among these 100 firms agreement is not perfect, but generally all five rating agencies share a common view. Companies such as Cisco, Nokia, and Colgate-Palmolive have high

[12] The average distance to the median across the 823 firms is 0.41, with the first quantile at 0.30 and the third quantile at 0.51

[13] The aggregate KLD rating has 14 unique values. These are the blue dots that seem to be aligned on top of each other.

median ratings, and all five rating agencies tend to agree. Firms such as Roper Industries, Intuitive Surgical, and China Resources Land, Ltd. have low median ratings, and all rating agencies agree with such an assessment. The average pairwise correlation of the ratings for these 100 firms is 0.90.

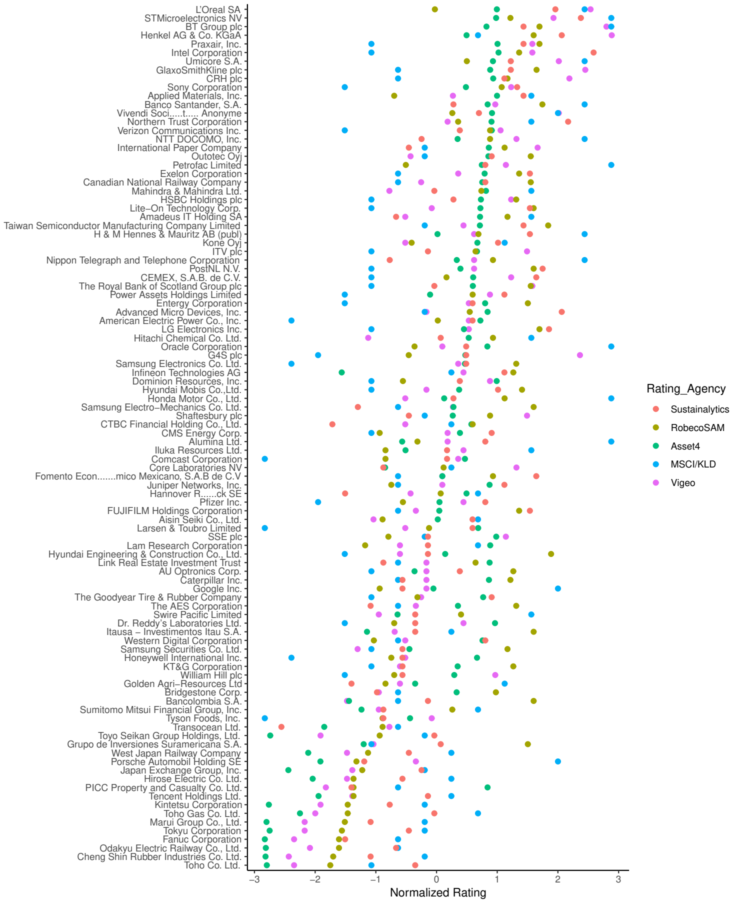


Figure 2. Comparison of firms' normalized scores for different rating agencies.

100 firms with the highest median average distance within the common sample (n=823). Firms within these group have been sorted by their respective median. Each rating agency ranking is plotted in a different color.

In Figure 2 we present a subset containing the 100 firms with the highest average distance to the median, i.e., where the disagreement between raters is greatest. It shows that there is variation across the spectrum. In the top 25 percentile of the median rating we can find firms such as Intel, GlaxoSmithKline, Applied Materials, and Sony. CEMEX, LG, Oracle, Samsung, Honda, Comcast, Pfizer, and Google are within the 50 and 75 percentile. Honeywell, Tyson Foods, Tencent, and Porsche are among the worst rated. Interestingly, independent of the rating of the firm, the disagreement in all of them is large. In fact, the average pairwise correlation of the ratings among this set of 100 firms is 0.32.

In summary, there is large heterogeneity in the level of disagreement across firms, measured both in correlations and average distance to the median. Rating agencies agree on some firms, and disagree on others. However, the level of disagreement does not seem to be related to the median level of the rating. For example, in Figure 2 there are firms with high scores and large disagreement, and in Figure 1 there are firms with low scores and large disagreement. L'Oreal and Nokia have very similar normalized median ratings, 1.96 and 1.43, respectively. Regarding L'Oreal the disagreement is on average 0.8 standard deviations from the median while regarding Nokia it is 0.23. Even though the median rating of L'Oreal is better, the disagreement is more than three times larger. Similar patters are found at every median rating.[14] Finally, disagreement occurs with smaller and bigger firms, and in all sectors and all countries in our sample.

3.3 Quantile Analysis

Although, Figures 1 and 2 show that there is discrepancy at all levels of the ratings, it is possible that the correlations and patterns of disagreement differ across different quantiles. Hence, rankings may be even more varied than correlations would imply.

The ranking can be more important than the individual score in many financial applications. Investors often want to construct a portfolio with sustainability leaders from the top quantile, or alternatively exclude sustainability laggards in the bottom quantile. With this approach, the disagreement on individual scores would be less relevant than the placement of the firm in comparison to its peers. To further evaluate this possibility we implemented a very simple procedure: We count how many firms are common across the five raters at the top and bottom 20 percent. The purpose is to evaluate if there is at least agreement on the firms belonging to the extremes of the distribution.

Table 3. Common set of firms among the top and bottom quantiles.

Common among Top quantile	**Common among Bottom quantile**
Akzo Nobel NV	Advance Auto Parts Inc.
Australia & New Zealand Banking Group Limited	Affiliated Managers Group Inc.
Aviva plc	America Movil S.A.B. de C.V.
BMW AG	Amphenol Corporation
BNP Paribas SA	Berkshire Hathaway Inc.
Campbell Soup Company	Cencosud S.A.
Commonwealth Bank of Australia	China Development Financial Holding Corporation
Dexus Property Group	China Resources Land Ltd.
Diageo plc	CP ALL Public Company Limited
EDP-Energias de Portugal, S.A.	Credit Saison Co. Ltd.
Hewlett-Packard Company	Deutsche Wohnen AG
Imperial Tobacco Group plc	Expedia Inc.
Industria de Diseno Textil SA	Genuine Parts Company
Kingfisher plc	Grupo Financiero Inbursa, S.A.B. de C.V.
Koninklijke Philips N.V	Hengan International Group Company Limited
Mondi plc	Intuitive Surgical, Inc.
National Australia Bank Limited	Japan Real Estate Investment Corporation
Nokia Corporation	Loews Corporation
Renault SA	MediaTek Inc.
Schneider Electric S.A.	MediPal Holdings Corporation
Solvay SA	Meiji Holdings Co., Ltd.
STMicroelectronics NV	Naver Corporation
Swiss Re Ltd	NCsoft Corporation
Telecom Italia S.p.A.	NEXON Co., Ltd.
UPM-Kymmene Oyj	Nippon Building Fund Inc.
Wipro Ltd.	Shimano Inc.
	Sumitomo Realty & Development Co. Ltd.
	Tokyo Electric Power Company, Incorporated

We calculate the intersection of the 164 best and worst rated firms from each rating agency, i.e., KLD, Sustainalytics, Vigeo-Eiris, RobecoSAM and Asset4 using the common sample of 823 firms in 2014.

Table 3 shows the number of common firms across all five raters for the top and bottom 20 percent of the firms in the common sample. The firms are sorted alphabetically within each group. The first column in Table 3 provides an idea of how a sustainable investment portfolio that is based on a consensus of five rating agencies would have looked like in 2014. There are only 28 firms that are

[14] For robustness, we computed the same Figures using the ranking as opposed to the normalized score. The results are even more striking when using rankings. The results are shown in Figures A.1 and A.2, in the appendix.

consistently in the bottom, and 26 that are consistently in the top. Most of the top rated companies are large and well-known companies. It is interesting that Diageo, Kingfisher, and Imperial Tobacco Group are among the companies that are consistently highly rated, given the health implications of their key products: alcohol and tobacco. A likely explanation is that rating agencies do not take into account the impact that firms have with their business model in their ESG performance assessment. For instance, for some raters, it does not make a difference for the ESG performance assessment whether a firm sells tobacco or a life saving drug as long as it does so in a sustainable way. The second column of Table 3 lists companies that one would expect to be consistently avoided by sustainable investment funds. We do not find any patterns regarding the size of firms or their industries except for one interesting observation, five of the 28 firms are domiciled in Japan.

In summary, there is large heterogeneity in the disagreement of the ranking of the firms and the results presented in Table 3 are sensitive to the size of the chosen quantile[15]. The disagreement on the rankings implies that the portfolio choice of the ESG top firms is strongly influenced by the choice of the rating agency. Furthermore, when investors base their decision on several rating agencies at once, there are only a few companies to choose from. At the same time, the small set of firms makes it very easy to claim that the worst performers are excluded, when only the consensus of different raters is considered.

To provide a more general description of the divergence, we devised a measure that we call the *Quantile Ranking Count*. First, we count how many common firms are in the lower $q\%$ of the common sample of all the rating agencies. We then calculate the ratio of this number to the total number of firms. If the rating agencies are perfectly aligned, then the exact same firms will be in the lower quantile ($q\%$). If the rating agencies completely disagree, then the probability that a firm is common to all rating agencies is q^n (n is the number of rating agencies) and the ratio of common firms over the sample size is small. Since we base our analysis on the common panel data, when the quantile is 100 percent, then all the firms are common to all the rating agencies and the ratio is exactly one. We denote this measure as the *Quantile Ranking Count* (QRC_q).

$$QRC_q = \frac{\text{Common Firms in the lower } q \text{ quantile}}{\text{Total Firms}} \tag{1}$$

In order to interpret the data, we simulated ratings with known and constant correlation. First, we simulated a random draw of 823 × 5 uniform realizations between the values of 0 and 1. Denote these realizations as $\epsilon_{k,f}$, where k is the rater and f is the index for the fictitious firm. Second, we created rankings for each rater and each firm as follows:

$$R_{kf} = \epsilon_{kf} + \alpha \times \sum_{x \neq k} \epsilon_{xf} \tag{2}$$

where the α is calibrated to achieve an average correlation across all ratings. A value of $\alpha = 0$ implies that all the ratings are perfectly uncorrelated, and $\alpha = 1$ implies perfect correlation. We calibrated the α to achieve an average correlation within sample of 10, 20,..., 80, 90, and 95 percent. Finally, from the simulated data we computed the Quantile Ranking Counts (QRC) for each quantile q.

In Figure 3 we present the Quantile Ranking Count for the overall ESG rating for all data providers and firms in the common sample. The plots for the environmental, social, and governance dimensions are shown in in the appendix in Figure A.3. The thick orange line indicates the counts of the actual

[15] See Appendix Figures A.1 and A.2 for the top and bottom 100 firms' disagreement.

data and the dashed gray lines reflect the implied counts of the simulated data. The quantiles range from five to 100 percent (x-axis) in increments of five percent. The implied correlations move from 0.1 to 1 in increments of 0.1 and are depicted in the gray lines. (we also added the 0.95 correlation simulation).

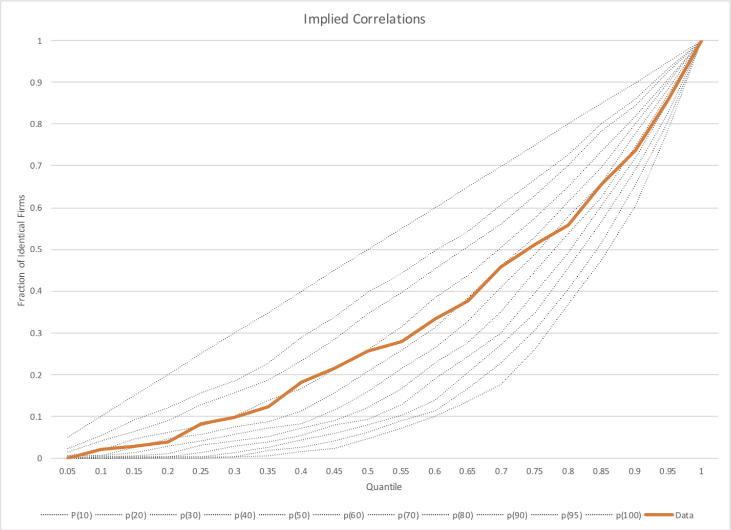


Figure 3. Quantile Ranking Count of ESG ratings including all rating agencies.

The gray lines represent simulated data, the implicit correlations, for each quantile from 10 to 100%. The orange line is the quantile ranking count for the true data, i.e., the fraction of identical companies in the sub sample of a given quantile.

First, let us concentrate on the 20 percent quantile to discuss the results. In Figure 3, the thick line is situated between the fifth and the sixth gray lines. This corresponds to an implied correlation between 60 and 70 percent. In other words, the implied correlation in the count of common firms among all the rating agencies is of the same order of magnitude as the one we would expect from data that is derived from rankings that are correlated between 60 and 70 percent. At the 50 percent quantile the thick line crosses the fourth gray line that corresponds to the 80 percent implied correlation. Finally, at the 90 percent quantile the implied correlation is 40 percent. This indicates that there is less agreement on the tails of the distribution than in the center. The lowest agreement is at the top end. Future research should explore the reasons behind this pattern.

The QRC documents the implied correlation at each quantile level; its curvature captures the overall implied correlation in rankings. We introduce a curvature measure, similar to the Gini coefficient, to evaluate the implied correlation of the QRC. As can be seen in Figure 3, an increase in the implied correlation decreases the curvature of the QRC. Our curvature measure can be understood as the inverse of the area that lies between the straight line that depicts perfect correlation and the line of the actual data counts. The higher the correlation, the higher the "Gini" coefficient. Figure 4 presents the curvature measures for the E, S , G and ESG ratings.

The environmental dimension has the highest implied correlation, followed by social and governance, respectively. ESG is situated between E and S. The implied correlation among E is between 0.7 and 0.8. The ESG ratings are just above 0.7, while the social and governance implied correlations are 0.65 and 0.57.

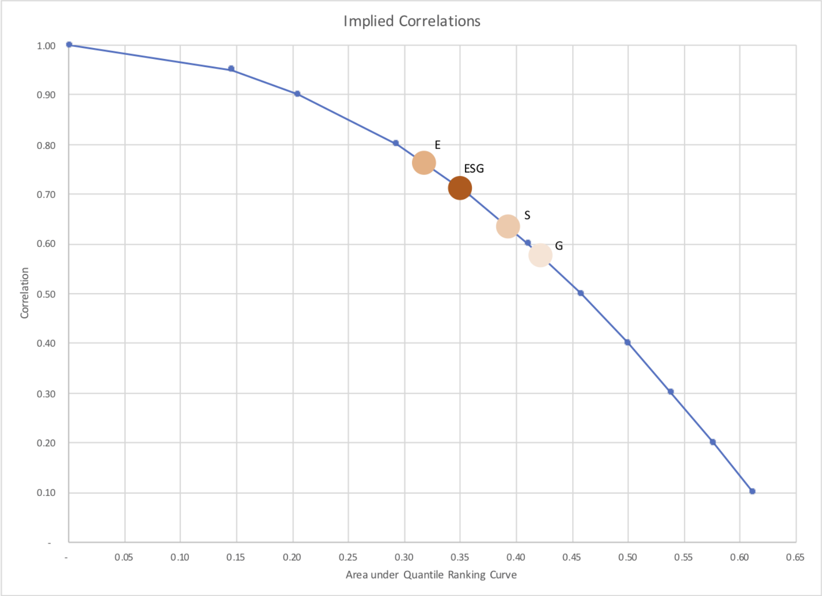


Figure 4. Gini coefficient for the Quantile Ranking Count for E,S,G and ESG for all Raters

The curvature measure (similar to the Gini coefficient) is used to evaluate the implied correlation in the Quantile Ranking Count (QRC).

In summary, in this subsection we have shown that agreement is stronger for the firms closer to the median, than it is for firms that are at the extremes of the distribution. Furthermore, the implied correlation using the Quantile Ranking Count is larger than the pairwise correlations of the individual ratings. At the ESG level, the individual ratings are correlated on average at 61 percent, while rankings are implicitly correlated at 70 percent. Lastly, we show that there is more agreement in the environmental dimension than in the social and governance dimensions. These stylized facts about rating divergence suggest that while there are clearly some commonalities between ESG ratings, they still disagree.

4 Taxonomy and Aggregation Rules

Environmental, social, and governance ratings are aggregate indices that can be described in terms of scope, indicators, and an aggregation function. Scope refers to the range of attributes that are considered to be part of ESG performance. For example, most rating agencies consider a firm's greenhouse gas emissions, but only some include electromagnetic radiation that a firm is emitting. Indicators correspond to the measurements of a given attribute, i.e., the kind of raw data that is used and how it is transformed into a numerical value. Even if raters agree on the attribute that should be measured, they might disagree on the way the attribute is measured. For example, if two raters want to measure discrimination against women, for instance, the first rater could look at the gender pay gap, whereas the other rater would use the percentage of women on the board and/or in the workforce. The two measures are very likely to be correlated but most likely deliver somewhat different results. Finally, the ratings are constructed through a function that transforms multiple indicators into one aggregate rating. These functions assign different weights per indicator, reflecting different preferences. A rating agency that is more concerned with carbon emissions than electromagnetic fields will assign different weights than a rating agency that cares equally about both issues. Furthermore, different industries might also have different weights as some attributes are judged more important to some industries than others.

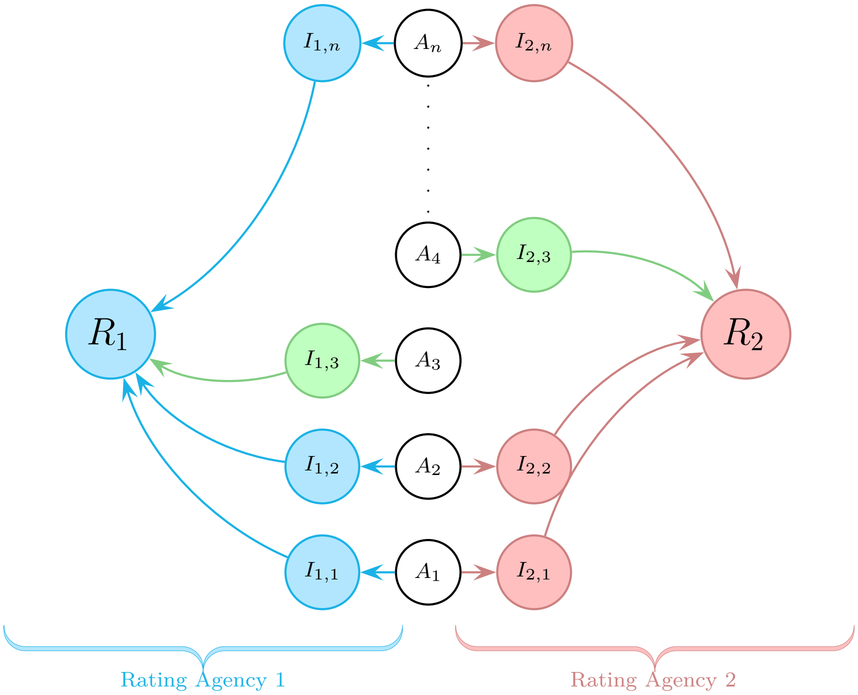


Figure 5. Rating Agencies Aggregation Procedures: Disentangling Discrepancies.

This general view of ESG ratings is illustrated in Figure 5. In the middle in white circles, there are n attributes denoted as "A", which represent all the attributes that can be thought of as relevant to ESG Performance. On the left and right, there are two different rating agencies, computing two different aggregate ratings R_1 and R_2. Divergence between these ratings can emerge from three distinct sources. The first source is *measurement*. Each attribute needs to be measured with an indicator, and the raters might use different indicators to do so. Figure 5 shows how each attribute is measured with rater-specific indicators, denoted as $I_{k,1}$, $I_{k,2}$, $I_{k,n}$, for rating agency k, in blue and red circles, respectively. The second source of divergence are differences in *scope*, i.e., the first rater chooses a different subset of attributes than the second rater. This situation is shown in Figure 5 in green, where rater 1 is the only one to consider Attribute A_3 and rater 2 is the only one to consider Attribute 4 A_4. Of course, if different attributes are considered then it is understandable that the overall rating will differ, too. The third source of divergence are differences in *weight*, shown by the arrows from the indicators to the rating. To progress from multiple indicators to one aggregate index, the raters need to use an aggregation function. This function could be an average, or a sum, but it could also be a more complex function involving nonlinear terms or contingencies on additional variables such as industry affiliation. Different aggregation functions will lead to different ratings, even if scope and measurement protocols are identical.

Technically, the divergence of scope could be subsumed under weight. The fact that a rating agency does not consider a particular attribute is equivalent to assuming that it sets the weight of that attribute to zero in the aggregation rule. Nevertheless, we believe it is informative to separate differences in scope from differences in weight. The measurement divergence, on the other hand, is purely a problem of using different indicators or proxies to try to quantify the same attribute.

4.1 Taxonomy

The goal of the paper is to decompose the overall divergence between ratings into the sources of measurement, scope, and weight. This is not trivial, because at the granular level, i.e., the most disaggregate data we have, the approach from each rating agency looks very different. Each rater chooses to break down the concept of ESG performance into different indicators, and presents those aspects in different hierarchies. For example, at the first level of disaggregation, Vigeo-Eiris, RobecoSAM, and Sustainalytics have three dimensions (E, S and G), Asset4 has four, and KLD has seven. Below these first level dimensions, there are between one and three levels of more granular sub-categories, depending on the rater. At the lowest level, our data set contains between 37 and 236 indicators per rater, which often, but not always, relate to similar underlying attributes. These diverse approaches make it difficult to understand how and why different raters assess the same company in different ways.

To develop the taxonomy of indicators we created a long list of all available indicators, including their detailed descriptions. In some cases, where the descriptions were not available (or were insufficient) we interviewed the data providers for clarification. We also preserved all additional information that we could obtain, such as to what higher dimension the indicator belongs or whether the indicator is industry-specific. In total, the list contained 641 indicators.

We define the taxonomy taking a bottom-up approach. First, we grouped similar indicators together, establishing common *categories* from the population of indicators. For example, we grouped together all indicators related to resource consumption or those related to community relationships. Next, we iteratively refined the taxonomy, following two rules. First, each indicator was assigned to only one category. Second, whenever at least two indicators from different raters both describe a category that is distinct from a previously existing category, they were combined in a new category. For example, indicators related to forests were taken out of the larger category of biodiversity to form their own category. Similarly, indicators related to reporting quality were taken out of various other existing categories to form their own category.

The taxonomy contains a total of 64 categories. Table 4 shows the categories, as well as how many indicators from each rater are sorted into each category. Some categories, such as GMOs (genetically modified organisms) contain just one indicator from two raters. Others, such as supply chain contain several indicators from all raters. The reason for this difference in the broadness of categories is that there were no indicators in supply chain that together represented a more detailed common category. Therefore, the comparison in the case of supply chain is at a more general level, and it may seem obvious that different raters take a different view of this category. Nevertheless, given the data, this broad comparison represents the most specific level possible. A total of 70 indicators remained unclassified. They are unique to one rater and could not be grouped with similar indicators from other raters. We assign these indicators to their own unique rater-specific category.

In our sample, Asset4 has the most individual indicators with 236, followed by Sustainalytics with 155. KLD and RobecoSAM have 75 and 74, respectively, and Vigeo-Eiris has 37. The zeros in Table 4 indicate that not all rating agencies cover all categories. This indicates differences in scope. There are zeros not only for categories that could be described as specialized, such as electromagnetic

Table 4. Number of indicators per Categories.

	KLD	Vigeo-Eiris	RobecoSAM	Sustainalytics	Asset4
Access to Basic Services	1	0	0	2	1
Access to Healthcare	1	0	3	6	1
Animal Welfare	0	0	0	2	1
Anti-Competitive Practices	1	1	0	0	2
Audit	0	1	0	4	7
Biodiversity	2	1	1	1	3
Board	0	1	0	6	26
Board Diversity	1	0	0	1	0
Board Gender Diversity	2	0	0	1	0
Business Ethics	1	0	2	4	1
Chairman Ceo Separation	0	0	0	1	1
Child Labor	1	1	0	0	1
Climate Risk Mgmt.	2	0	2	0	1
Clinical Trials	0	0	0	1	1
Collective Bargaining	0	1	0	2	1
Community and Society	1	1	6	3	10
Corruption	1	1	0	2	1
Customer Relationship	2	1	1	1	7
Discrimination and Diversity	3	1	0	2	9
ESG incentives	0	0	1	1	0
Electromagnetic Fields	0	0	1	1	0
Employee Development	3	1	2	1	13
Employee Turnover	0	0	0	1	1
Energy	1	1	6	3	5
Environmental Fines	1	0	0	1	1
Environmental Mgmt. System	1	0	0	2	1
Environmental Policy	0	2	3	4	4
Environmental Reporting	0	0	1	2	1
Financial Inclusion	1	0	0	1	0
Forests	0	0	1	1	0
GHG Emissions	1	1	0	5	5
GHG Policies	0	0	2	3	4
GMOs	0	0	1	1	1
Global Compact Membership	0	0	0	1	1
Green Buildings	1	0	2	5	1
Green Products	1	1	1	7	20
HIV Programmes	0	0	0	1	1
Hazardous Waste	0	0	1	1	1
Health and Safety	2	1	1	7	7
Human Rights	5	1	1	2	5
Indigenous Rights	1	0	0	1	1
Labor Practices	3	4	1	3	16
Lobbying	0	1	1	3	0
Non-GHG Air emissions	0	0	0	1	2
Ozone Depleting Gases	0	0	0	1	1
Packaging	1	0	1	0	0
Philanthrophy	1	1	1	3	2
Privacy and IT	2	0	3	1	0
Product Safety	6	3	2	2	13
Public Health	2	0	3	1	0
Remuneration	4	2	1	4	15
Reporting Quality	1	0	0	3	5
Resource Efficiency	0	0	3	1	6
Responsible Marketing	1	1	3	3	1
Shareholders	0	1	0	0	16
Site Closure	0	0	1	1	0
Supply Chain	6	4	3	21	4
Sustainable Finance	4	0	5	9	3
Systemic Risk	1	0	1	0	0
Taxes	0	0	1	2	1
Toxic Spills	1	0	0	1	2
Unions	1	0	0	0	1
Waste	3	1	2	3	4
Water	2	1	2	2	3
Unclassfied	2	1	7	7	40
Sum	78	38	80	163	282

We consider a category to be covered when at least one firm is rated within that category for a given rating agency. This is a very low threshold.

radiation, but also for the category taxes, which may seem like a fundamental concern in the context of ESG. Also, the considerable number of unclassified indicators shows that there are many aspects of ESG that are only measured by one out of five raters. Most of the unclassified indicators stem from Asset4's economic dimension, which is not covered in any other rating agency. However, there are some unclassified indicators from each rater.

The common aspects that are considered in all five ratings are community and society, customer relationship, employee development, energy, green products, health and safety, labor practices, product safety, remuneration, responsible marketing, supply chain, waste, and water. There are also 17 matches that are explicitly considered only by two rating agencies, namely animal welfare, chairman/CEO separation, child labor, clinical trials, electromagnetic fields, employee turnover, environmental fines, financial inclusion, global compact membership, HIV programs, lobbying, non-GHG air emissions, ozone-depleting gases, reporting quality, shareholders, site closure, and systemic risk.

The taxonomy allows comparing the ratings at the level of categories. To do so, we created category scores (C) for each category, firm, and rater. Category scores were calculated by taking the average of the indicator values assigned to the category. Let us define the notations:

Definition 1 ***Category Scores, Variables and Indexes:***
The following variables and indexes are going to be used throughout the paper:

Notation	Variable	Index	Range
A	Attributes	i	$(1, n)$
I	Indicators	i	$(1, n)$
C	Categories	j	$(1, m)$
N_{fkj}	Indicators $\in C_{fkj}$	i	$(1, n_{fkj})$
R	Raters	k	$(1, 5)$
F	Firms	f	$(1, 823)$

The category score is computed as:

$$C_{fkj} = \frac{1}{n_{fkj}} \sum_{i \in N_{fkj}} I_{fki} \tag{3}$$

for firm f, rating agency k, and category j.

Category scores represent a rating agency's assessment of a certain ESG category. They are based on different sets of indicators that each rely on different measurement protocols. It follows that differences between category scores stem from differences in *how* rating agencies choose to measure, rather than what they choose to measure. Thus differences between the same categories from different raters can be interpreted as measurement divergence. Furthermore, rating agencies may employ different sets of indicators depending on the firms' industries. Therefore, the category scores may consist of a different set of indicators for different firms even for the same rating agency. In our procedure, the different views at this level of granularity will be measured as disagreement about measurement instead of scope. This also implies that our linear estimations in the following sections are allowing for sectoral differences in so far that the average measure within each category captures the industry specific indicators.

Table 5 shows the correlations between the categories. The correlations are calculated on the basis of complete pairwise observations per category and rater pair[16]. They range from -0.47 for responsible marketing between KLD and Sustainalytics to 0.81 for remuneration between Sustainalytics and Asset4. When comparing the different rater pairs, Vigeo-Eiris and RobecoSam have the highest average correlation with 0.47, and the pairs including KLD have all relatively low correlations ranging from 0.12 to 0.21.

Table 5. Correlation between rating agencies at the level of categories.

	KL:A4	**KL:RS**	**KL:SA**	**KL:VI**	**RS:A4**	**RS:SA**	**SA:A4**	**VI:A4**	**VI:RS**	**VI:SA**	**Average**
Access to Basic Services	0.34		0.42				0.48				0.41
Access to Healthcare	0.52	0.53	0.59		0.55	0.65	0.65				0.58
Animal Welfare							0.61				0.61
Anti-Competitive Practices	0.55			-0.04				-0.05			0.15
Audit							0.64	0.57		0.55	0.58
Biodiversity	0.03	0.00		0.04	0.47			0.41	0.67		0.27
Board							0.51	0.59		0.38	0.49
Business Ethics	0.34	-0.07	0.05		-0.11	0.34	-0.03				0.09
Chairman Ceo Separation							0.56				0.56
Child Labor	0.49										0.49
Climate Risk Mgmt.	0.44	0.45			0.56						0.48
Clinical Trials							0.60				0.60
Collective Bargaining							-0.05	0.00		0.51	0.16
Community and Society	0.20	0.21	-0.24	0.28	0.58	-0.15	-0.01	0.50	0.50	-0.07	0.18
Corruption	-0.13		0.27	0.30			-0.18	-0.16		0.56	0.11
Customer Relationship	-0.04	-0.08	0.27	-0.08	0.46	-0.17	-0.15	0.52	0.50	-0.08	0.12
Discrimination and Diversity	0.00		-0.04	-0.03			0.61	0.62		0.63	0.30
Electromagnetic Fields						0.49					0.49
Employee Development	0.34	0.32	0.00	0.29	0.57	0.18	0.32	0.29	0.38	0.19	0.29
Employee Turnover							0.40				0.40
Energy	0.27	0.31	0.12	0.24	0.22	0.26	0.30	0.45	0.37	0.38	0.29
Environmental Fines							0.17				0.17
Environmental Mgmt. System	-0.07		0.63				0.41				0.32
Environmental Policy					0.63	0.50	0.52	0.62	0.63	0.52	0.57
Environmental Reporting					0.39	0.52	0.25				0.39
Financial Inclusion			0.29								0.29
Forests											
GHG Emissions	-0.17		-0.11	-0.05			0.35	0.48		0.30	0.13
GHG Policies					0.41	0.28	0.68				0.45
GMOs					0.19	0.44	0.41				0.35
Global Compact Membership							0.86				0.86
Green Buildings	0.22	0.48	0.56		0.08	0.39	0.21				0.32
Green Products	0.38	0.28	0.26	0.13	0.56	0.40	0.52	0.35	0.38	0.46	0.37
HIV Programmes							0.73				0.73
Hazardous Waste						0.20	0.09				0.15
Health and Safety	0.28	0.24	0.04	0.30	0.58	-0.15	-0.17	0.71	0.63	-0.14	0.23
Human Rights	0.11		-0.10	0.13			0.05	0.46		0.01	0.11
Indigenous Rights	-0.10		0.35				-0.27				-0.01
Labor Practices	0.05	-0.13	0.13	-0.03	0.38	0.18	0.34	0.48	0.55	0.19	0.21
Lobbying										-0.28	-0.28
Non-GHG Air emissions							0.42				0.42
Ozone Depleting Gases							0.62				0.62
Packaging											
Philanthropy					0.26	0.39	0.43	0.28	0.42	0.43	0.37
Privacy and IT		0.32	0.36			0.27					0.32
Product Safety	0.02	0.19	0.02	0.05	0.37	-0.10	-0.05	0.25	0.49	-0.09	0.11
Public Health		0.49	0.46			0.47					0.47
Remuneration	0.13	0.00	0.14	0.08	0.29	0.17	0.81	0.73	0.19	0.69	0.32
Reporting Quality							0.51				0.51
Resource Efficiency					0.59	0.33	0.34				0.42
Responsible Marketing	0.20	-0.34	-0.47	-0.08	-0.11	0.60	-0.07	0.00	0.43	0.40	0.06
Shareholders								0.43			0.43
Site Closure						0.34					0.34
Supply Chain	0.16	0.11	0.17	0.17	0.56	0.53	0.53	0.63	0.64	0.56	0.41
Sustainable Finance	0.49	0.45	0.58		0.63	0.67	0.69				0.59
Systemic Risk		0.26									0.26
Taxes					-0.03	0.11	0.00				0.03
Toxic Spills	0.03		-0.21				0.07				-0.04
Unions	0.66										0.66
Waste	0.27			0.33	0.23			0.38	0.28		0.30
Water	0.23	0.20	0.31	0.32	0.12	0.42	0.40	0.40	0.47	0.47	0.33
Average	0.21	0.20	0.18	0.12	0.36	0.31	0.34	0.40	0.47	0.30	

Correlations between the different categories from different rating agencies. We calculate a value for each criterion on the firm level by taking the average of the available indicators for firm f and rater k. The panel is unbalanced due to differences in scope of different ratings agencies and categories being conditional on industries. SA, RS, VI, A4, and KL are short for Sustainalytics, RobecoSAM, Vigeo-Eiris, Asset4, and KLD, respectively.

Beyond these descriptive observations, Table 5 offers two insights. First, the average level of correlations between categories is markedly lower than the correlations between the aggregate ratings as reported in Table 2. For example, the correlations of the categories water and energy with an average of 0.33 and 0.29, respectively, are much lower than of the environmental dimension with an average of 0.70. Hence, the divergence increases with granularity. This finding is surprising because

[16] Table A.1 in the appendix shows the number of complete observations that lie at the basis of Table 5.

we would have expected less disagreement on specific category scores, and more disagreement at the aggregate level. This is because the aggregate rating is affected by differences in scope and aggregation rule, whereas category scores should only be affected by measurement divergence. Future research should study the reasons behind the disagreement at different levels of aggregation. This is beyond the scope of the current paper.

The second insight is that there are large differences in terms of correlation levels. Environmental policy, for instance, has an average correlation level of 0.57. This indicates that there is at least some level of agreement regarding the existence and quality of the firms' environmental policy. However, most categories exhibit lower correlations. Surprisingly, even categories that measure straightforward facts that are easily obtained from public records have very heterogeneous levels of correlation. Membership in the UN Global Compact and CEO/Chairman separation, for instance, show correlations of 0.86 and 0.56, respectively. There are also a number of negative correlations. They appear mostly in categories of the social dimension, such as responsible marketing and occupational health and safety, but also in the category toxic spills. This indicates that the level of disagreement is so severe on some categories that rating agencies reach not just different, but even opposite conclusions.

Imposing a taxonomy on the pool of indicators, i.e. classifying indicators into categories, requires some subjective judgment. To limit the effect of subjective bias, the classification was proposed by one author, and then audited by another author, and each case of disagreement was discussed and resolved. To make sure our results are not driven by a particular classification, we created an alternative taxonomy. Instead of constructing the categories from the bottom up, we produced a top-down taxonomy that relies on external categories established by the Sustainability Accounting Standards Board. In a comprehensive stakeholder consultation process, SASB has identified 26 so-called 'general issue categories'[17]. We mapped all indicators against these 26 general issue categories, forcing each indicator to be assigned only to one category. The results are quite similar with the alternative taxonomy[18].

The taxonomy and the category scores reveal that there are large differences in scope, i.e., not all raters cover all categories of the taxonomy, as well as striking differences in measurement, i.e., the categories are only weakly correlated. The results presented are robust to changes in the taxonomy, even when using the more aggregated SASB categories, where scope should be lower and measurement exacerbated[19].

4.2 Aggregation Rule Estimation

In this subsection we estimate the aggregation rule used by the rating agencies. Our purpose is to "reverse-engineer" the function that aggregates the category scores (C_{fkj}) to the rating (R_{fk}) for rater k. The procedure is described in Figure 6, where the category scores are intermediate indexes in the computation of the aggregation rule. We will use these results as a basis to decompose the disagreements between raters in scope, measurement, and weight divergence and we need comparable functions across rating agencies to do so. We use the R^2 within sample as our measure of quality of fit given that our objective is in-sample predictability.

[17] https://materiality.sasb.org

[18] We refer the reader to Appendix A for the detailed results that are based on the alternative taxonomy.

[19] See table A.3 for the numbers of indicators and table A.4 for the correlations using the SASB taxonomy in the appendix.

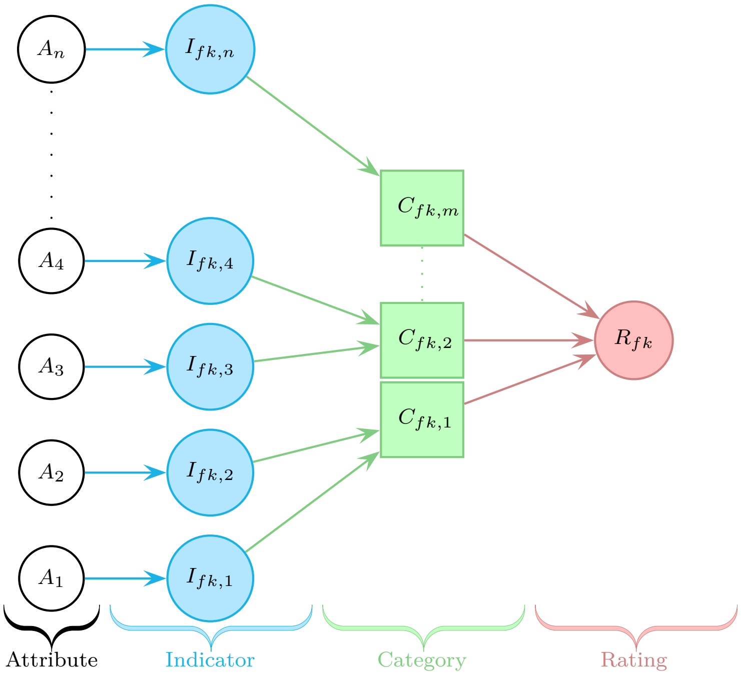


Figure 6. Estimation Procedure for the Aggregation Rule (Rating Agency k, Firm f)

The analysis relies on the category scores and the taxonomy established in chapter 4.1. We compute category scores for the common sample according to our taxonomy and also include each unclassified indicators as separate rater-specific categories. When there are no indicator values available to compute the category score for a given firm, the score is set to zero [20]. Finally, we drop categories altogether when there are no values available for any firm in the common sample.

Our preferred specification is a simple linear regression with sign restrictions on the coefficients (coefficients need to be non-negative). We estimate the weights (w_{kj}) with the following specification:

$$\begin{aligned} R_{fk} &= \sum_{j\in(1,m)} C_{fkj} \times w_{kj} + \epsilon_{fk} \\ w_{kj} &\geq 0 \end{aligned}$$

As the data was previously normalized, we exclude the constant term. Due to the non-negative constraint we calculate the standard errors by bootstrap. For a given rater, categories that do not exist or do not contain any data are marked with dashes. The results are shown in Table 6. The two lowest R^2 are 0.90 for Sustainalytics and 0.92 for Asset4. The regressions for KLD, Vigeo-Eiris, and RobecoSAM have R^2 of 0.99, 0.96, and 0.98, respectively. The high R^2 indicate that a linear model based on our taxonomy is able to replicate the original ratings quite accurately[21].

We evaluated several other possibilities: simple linear regression, random forests, and neural networks (with one layer or multiple layers). None of these specifications resulted in major improvements

[20] This is necessary in order to run regressions without dropping all categories with missing values. Of course, it entails an assumption that no information is a proxy for poor performance in a category. This assumption, however, does not seem to have a strong influence on the quality of fit. As we will show later, random forest regressions, which offer a way to relax this assumption, do not yield large improvements.

[21] The coefficients in these regressions should not add to one for several reasons, the most simple one is that the variables have been normalized at the category level.

over the non-negative linear regression. When estimating the unrestricted linear models even though the R^2's fluctuated, they only changed by a maximum of 0.01[22]. When the estimation was performed allowing for a non-linear and flexible functional form, the improvements were very small. Estimating random forests produces R^2 of 0.93, 0.98, 0.99, 0.95, and 0.98 for KLD, Vigeo-Eiris, RobecoSAM, Sustainalytics, and Asset4, respectively. Estimating a two-layer neural network with a linear activation function yields 0.98, 0.98, 0.99, 0.93, and 0.96, respectively. We also tried non-linear activation functions such as *relu* and *sigmoid*. In this case, the results deteriorated and the maximum R^2 value was 0.57. The results are also robust to using a different year and the full sample[23].

In another robustness check, we evaluated the fit of the regression assigning randomly 10 percent of the firms to a testing set, and the rest to a training set. The out-of-sample R^2 for KLD, Vigeo-Eiris, RobecoSAM, Sustainalytics, and Asset4 are 0.99, 0.94, 0.98, 0.88, and 0.86, respectively. The explanatory power in the out-of-sample is very close to the in-sample. The best fit is KLD. Asset4 is the one that performs the worst with a very reasonable decline of less than 6 percent in the R^2. As aggregation rules are subject to change through time, we do not run tests where the in-sample belongs to a different year than the out-of-sample.

We also replicated the estimation of the aggregation rule using the SASB taxonomy. The regressions using the same non-negative constraints produce R^2 of 0.98, 0.96, 0.98, 0.90, and 0.92 for KLD, Vigeo-Eiris, RobecoSAM, Sustainalytics, and Asset4, respectively[24]. The results are virtually identical reflecting that the classification of the categories has a small impact on the overall fit. This is due, in part, to the indicators that are unclassified in both taxonomies. Another reason is the rater effect discussed in subsection 5.2. When the rater effect is high, the marginal explanatory power of each additional category is diminishing, i.e., within rater categories are correlated with each other.

In our last robustness check, we run ordinary least square regressions of ratings on indicators to see if the fit improves and to evaluate how much of the fit is lost by the categorization. The R^2 are 1, 0.96, 0.99, 0.93, and 0.95 for KLD, Vigeo-Eiris, RobecoSAM, Sustainalytics, and Asset4, respectively. The biggest changes are Asset4 and Sustainalytics. They both go up by 0.03. Overall, the changes are minor.

The regression coefficients represent an explicit approximation of each rater's aggregation rule. These estimated aggregation rules can now be compared to determine the relative importance of each of the categories. In other words, our coefficients' estimate the implied tradeoffs between categories[25]. There are substantial differences in the weights for different raters. The three most important categories for KLD are climate risk management, product safety, and remuneration. For Vigeo-Eiris, they include discrimination and diversity, environmental policy, and labour practices. For RobecoSAM they are employee development, climate risk management, and resource efficiency. Sustainalytics ranks supply chain, green products, and environmental management system as its three most important. For Asset4, board, resource efficiency, and remuneration are their three most important. Only resource efficiency and climate risk management are among the three most substantial categories for more than one rater, showing that different raters have strongly diverging views about which categories are most relevant. Furthermore, there are categories that have zero weight for all raters, such as board diversity and environmental fines, GMOs, HIV programs, ozone-depleting gases, and site closure.

[22] See Table A.2 in the appendix

[23] Results are available upon request.

[24] See Table A.5 in the appendix for the coefficients.

[25] Future research should explore the appropriateness of the weights in the aggregation rule.

Table 6. Estimates of Non Negative Least Squares Regression.

	KLD	Vigeo-Eiris	RobecoSAM	Sustainalytics	Asset4
Access to Basic Services	0.073***	-	-	0.016	0.000
Access to Healthcare	0.047***	-	0.006	0.042***	0.000
Animal Welfare	-	-	-	0.056***	0.000
Anti-Competitive Practices	0.130***	0.021**	-	-	0.058***
Audit	0.000	0.088***	-	0.000	0.019
Biodiversity	0.069***	0.019**	0.000	0.000	0.000
Board		0.116***	-	0.064***	0.188***
Board Diversity	0.000	-	-	0.000	0.000
Board Gender Diversity	0.000	-	-	0.051***	-
Business Ethics	0.147***	-	0.051***	0.099***	0.002
Chairman Ceo Separation	-	-	-	0.043***	0.009
Child Labor	0.049***	0.000	-	-	0.006
Climate Risk Mgmt.	0.232***	-	0.136***	-	0.064***
Clinical Trials	-	-	-	0.000	0.000
Collective Bargaining	-	0.061***	-	0.055***	0.010
Community and Society	0.129***	0.004	0.083***	0.082***	0.026
Corruption	0.122***	0.075***	-	0.051***	0.014
Customer Relationship	0.099***	0.030***	0.098***	0.113***	0.085***
Discrimination and Diversity	0.026**	0.153***	-	0.094***	0.058***
ESG incentives	-	-	0.000	0.017	-
Electromagnetic Fields	-	-	0.000	0.031**	-
Employee Development	0.144***	0.067***	0.223***	0.016	0.107***
Employee Turnover	-	-	-	0.008	0.005
Energy	0.050***	0.105***	0.014**	0.025*	0.031**
Environmental Fines	0.000	-	-	0.000	0.000
Environmental Mgmt. System	0.212***	-	-	0.181***	0.000
Environmental Policy	-	0.170***	0.098***	0.101***	0.009
Environmental Reporting	-	-	0.040***	0.047***	0.005
Financial Inclusion	0.060***	-	-	0.000	-
Forests	-	-	0.017**	0.012	-
GHG Emissions	0.031***	0.040***	-	0.046**	0.000
GHG Policies	-	-	0.009**	0.100***	0.029
GMOs	-	-	0.000	0.000	0.000
Global Compact Membership	-	-	-	0.027**	0.000
Green Buildings	0.072***	-	0.063***	0.069***	0.000
Green Products	0.130***	0.024***	0.034***	0.162***	0.101***
HIV Programmes	-	-	-	0.000	0.000
Hazardous Waste	-	-	0.000	0.022*	0.000
Health and Safety	0.173***	0.125***	0.043***	0.062***	0.058***
Human Rights	0.140***	0.000	0.000	0.066***	0.075***
Indigenous Rights	0.095***	-	-	0.028	0.000
Labor Practices	0.130***	0.145***	0.058***	0.000	0.068***
Lobbying	-	0.016*	0.000	0.091***	-
Non-GHG Air emissions	-	-	-	0.014	0.000
Ozone Depleting Gases	-	-	-	0.000	0.000
Packaging	0.047***	-	0.000	-	-
Philantrophy	0.000	0.071***	0.077***	0.032*	0.041***
Privacy and IT	0.124***	-	0.040***	0.028**	-
Product Safety	0.225***	0.065***	0.002	0.026**	0.052***
Public Health	0.080***	-	0.011	0.000	-
Remuneration	0.222***	0.108***	0.039***	0.000	0.127***
Reporting Quality	0.000	-	-	0.135***	0.097***
Resource Efficiency	0.000	0.000	0.116***	0.003	0.135***
Responsible Marketing	0.077***	0.006	0.025***	0.000	0.000
Shareholders	-	0.094	-	-	0.103***
Site Closure	-	-	0.000	0.006	-
Supply Chain	0.132***	0.060***	0.055***	0.248***	0.040***
Sustainable Finance	0.090***	-	0.071***	0.064***	0.050***
Systemic Risk	0.100***	-	0.049***	-	-
Taxes	-	-	0.008	0.040***	0.026**
Toxic Spills	0.113***	-	-	0.000	0.019
Unions	0.155***	-	-	-	0.012
Waste	0.195***	0.013	0.007	0.000	0.031**
Water	0.177***	0.000	0.020***	0.038***	0.031**
Unclassified Indicators	Yes	Yes	Yes	Yes	Yes
R2	0.99	0.96	0.98	0.90	0.92
Observations	823	823	823	823	823

Non-negative linear regressions of the most aggregate rating (ESG) on the categories of the same rater. As categories depend on industries we fill missing values of the independent variables with zeros. ***,** and * denote statistical significance at the one, five and ten percent level, respectively. As the data was previously normalized, we exclude the constant term. The standard errors are bootstrapped. Non-existent categories are denoted as dashes.

5 Decomposition and Rater Effect

In this section, we use the estimates to first decompose the differences between ratings in three components: scope, measurement, and weight. We then evaluate the patterns behind the measurement disagreements. We find that the differences are highly correlated within rating agencies, namely we detect a rater effect.

5.1 Scope, Measurement and Weight Divergence

We developed two alternative approaches for the decomposition. First, we arithmetically decompose the difference between two ratings into contributions in scope, measurement, and weight. Second, we explain one rating with scope, measurement, and weight variables that we construct using the information of another rater. The advantage of the first procedure is that we are directly decomposing the differences between two raters. However, the correlation between the different measures of divergence makes the individual contribution of each hard to disentangle. The second procedure is a variance decomposition that allows us to control partially for the correlations between the different measures of divergence. However, it does not allow us to look at the exact differences, and it only yields upper and lower bounds for each source of divergence.

5.1.1 Arithmetic Decomposition

The arithmetic variance decomposition assumes that all ratings are linear combinations of their categories. This assumption is reasonable based on the quality of fit of the linear estimations from section 4.2. With this assumption in place, we can explicitly calculate how scope, measurement, and weight divergence contribute to the overall difference between two ratings. The intuition is that the difference due to scope can be separated by looking only at the categories that are exclusively contained in one of the two ratings. The differences due to measurement can be isolated by calculating both ratings with the common categories and a set of weights common to both raters, so that differences can only stem from differences in measurement. The weight divergence is what remains of the total difference.

Let R_{fk} (where $k \in a, b$) be vectors of ratings provided by rating agency a and rating agency b for a common set of f companies. The ratings R_{fk} are represented by the vector product of category scores C_{fkj} and rating agency specific weights w_k, plus an error term that represents the difference between the true rating and the fitted rating. $\hat{R}_{fk}$ denotes the fitted rating and $\hat{w}_{kj}$ the estimated weight for rater k and category j. Our estimation of the aggregation rule is therefore:

$$\hat{R}_{fk} = C_{fkj} \times \hat{w}_{kj} + \epsilon_{fk}$$

Some categories are common to both raters, denoted as $C_{fkj_{com}}$. Other categories are exclusively measured by each rater, denoted as $C_{faj_{a,ex}}$ and $C_{fbj_{b,ex}}$, where $j_{a,ex}$ ($j_{b,ex}$) is the set of categories that are measured by rating agency a but not b (b but not a). Furthermore, $\hat{w}_{aj_{a,ex}}$ are the weights for the categories only measured by a, $\hat{w}_{bj_{b,ex}}$ for b. $\hat{w}_{aj_{com}}$, and $\hat{w}_{bj_{com}}$ are the weights by rating agencies a and b in the common categories. We define the fitted ratings for the common and exclusive sets of categories as follows:

Definition 2 *Fitted Rating in common and exclusive categories. For $k \in \{a, b\}$ define:*

$$\begin{aligned}
\hat{R}_{fk,com} &= C_{fkj_{com}} \times \hat{w}_{kj_{com}} \\
\hat{R}_{fk,ex} &= C_{fkj_{k,ex}} \times \hat{w}_{kj_{k,ex}} \\
\hat{R}_{fk} &= \hat{R}_{fk,com} + \hat{R}_{fk,ex}
\end{aligned}$$

$\hat{R}_{fk,com}$ is the fitted rating calculated with the common categories of rater k, $\hat{R}_{fk,ex}$ is the fitted rating calculated with the exclusive categories of rater k and $\hat{R}_{fk}$ is the sum of the two, i.e., the fitted ESG rating.

We summarize this discussion and the definition of the scope, measurement, and weight variables in the following definition:

Definition 3 ***Scope, Measurement, and Weight Variables***

$$\begin{aligned}
\Delta_{scope} &= \hat{R}_{fa,ex} - \hat{R}_{fb,ex} &&= (C_{faj_{a,ex}} \times \hat{w}_{aj_{a,ex}} - C_{fbj_{b,ex}} \times \hat{w}_{bj_{b,ex}}) \\
\Delta_{meas} & &&= (C_{faj_{com}} - C_{fbj_{com}}) \times \hat{w}^* \\
\Delta_{weight} &= \hat{R}_{fa,com} - \hat{R}_{fb,com} - \Delta_{meas} &&= (C_{faj_{com}} \times \hat{w}_{aj_{com}} - C_{fbj_{com}} \times \hat{w}_{bj_{com}}) - \Delta_{meas}
\end{aligned} \tag{4}$$

where $\hat{w}^*$ are the estimates from pooling regressions using the comon categories from rater a and b.

$$\begin{pmatrix} \hat{R}_{fa,com} \\ \hat{R}_{fb,com} \end{pmatrix} = \begin{pmatrix} C_{faj_{com}} \\ C_{fbj_{com}} \end{pmatrix} \times w^* + \begin{pmatrix} \epsilon_{fa} \\ \epsilon_{fb} \end{pmatrix} \tag{5}$$

We are interested in the difference between the ratings $\Delta_{a,b}$, which can be decomposed into three components:

$$\Delta_{fa,b} = \hat{R}_{fa} - \hat{R}_{fb} = \Delta_{scope} + \Delta_{meas} + \Delta_{weights} \tag{6}$$

This decomposition can be derived from the linear aggregation rules.[26] The intuition of the decomposition is as follows: Scope is captured by the difference in the fitted rating that is calculated using only the exclusive categories. We denote this Δ_{scope}. Second, to determine the contribution of measurement we evaluate the difference in fitted ratings that are calculated based on the common categories and the same aggregation weights for both raters. Equation 5 is a linear pooling regression of the fitted ratings of the two raters on the common categories of the two raters.[27] This way we restrict the weights to be the same across the two rating agencies.[28] Since the ordinary least squares make sure that we maximize the fit with $\hat{w}^*$, we can deduce that Δ_{meas} captures the differences that are exclusively coming from differences in the category scores. Finally, the contribution of Weight ($\Delta_{weights}$) is computed as the residual of the difference between the fitted ratings based on the common categories minus the measurement divergence from the previous step. The sum of these three components is an exact decomposition of the disagreement of the fitted values of the two rating agencies.

[26] In the non-linear estimation case, the decomposition can still be computed, but the interpretations are very different. Future research should study the robustness of the results presented here to a non-linear rule. In our case, the fit within sample is high enough that this decomposition is a very good approximation.

[27] We stack the fitted ratings of the two raters on each other in a single vector. The common categories of the two raters are stacked on each other in a single firm-by-categories matrix. We then regress the vector on the matrix using ordinary least squares.

[28] Of course, the quantitative results change when different weights are used, but in our case the qualitative results remained unchanged.

Table 7. Arithmetic Decomposition.

		Scope	Measurement	Weights	Residuals	Fitted	True
KLD	**Vigeo-Eiris**	0.42	0.61	0.27	0.17	0.79	0.80
KLD	**RobecoSam**	0.35	0.62	0.32	0.13	0.79	0.80
KLD	**Sustainalytics**	0.32	0.55	0.31	0.26	0.73	0.77
KLD	**Asset4**	0.35	0.58	0.47	0.25	0.80	0.87
Vigeo-Eiris	**RobecoSam**	0.32	0.38	0.11	0.17	0.61	0.62
Vigeo-Eiris	**Sustainalytics**	0.39	0.51	0.24	0.28	0.54	0.60
Vigeo-Eiris	**Asset4**	0.30	0.48	0.18	0.28	0.54	0.62
RobecoSam	**Sustainalytics**	0.32	0.54	0.17	0.26	0.59	0.65
RobecoSam	**Asset4**	0.27	0.50	0.16	0.26	0.62	0.71
Sustainalytics	**Asset4**	0.18	0.45	0.33	0.32	0.54	0.65
Average		0.32	0.52	0.26	0.24	0.66	0.71

Results from the arithmetic decomposition. First, we estimate the weights by regressing the ESG rating of rater a on the categories of the same rater. We do the same for rater b. Second, we construct two different ratings for rater a and b by only taking mutually exclusive categories and using the weights from step 1. The mean absolute deviation of the differences of those two ratings reflects the differences in scope between the two rating agencies. Third, we stack the two firm-by-categories matrices of the common categories as well as the two fitted ratings for the common categories of rater a and b on each other and calculate a new set of weights that is thus common to both raters using ordinary least squares. We then subtract the newly fitted ratings based on the common weights of rater b from rater a and calculate the mean absolute deviation to determine the divergence in measurement. Fourth, we calculate the divergence stemming from the aggregation weight by subtracting the residuals from the previous step of rater b from rater a and calculate the mean absolute deviation. The column "Residuals" reports the mean absolute deviation of the differences of the residuals of two respective regressions ESG scores on categories, the column "Fitted" shows the the mean absolute deviation of the differences of the fitted values corresponding to the residuals of the previous column and "True" the actual ESG scores.

The results are presented in Table 7. The first three columns represent the decomposition between scope, measurement, and weight. The last three columns highlight the quality of the fit for illustrative purposes. We report the mean absolute deviation for each source of divergence. Since the ratings have been normalized to have mean zero and variance one, the mean absolute difference can be understood as a measure in terms of standard deviations. The analysis reveals that on average across all rater pairs, measurement divergence is 0.52 standard deviations, ranging from 0.38 to 0.62. Scope divergence causes an average shift of 0.32 standard deviations, ranging from 0.18 to 0.42. Weight divergence causes an average shift of 0.26 standard deviations, ranging from 0.11 to 0.47. While all sources of divergence are important, measurement divergence stands out as the most influential source.

The last column "True" compares the actual ratings from the rating agencies: $|R_{fk_1} - R_{fk_2}|$. The table shows that the degree of discrepancies are of the order of 0.70 standard deviations. The biggest discrepancies are between KLD and the other raters. The strongest agreement is between Vigeo-Eiris and Sustainalytics. The column "Fitted" shows the mean absolute difference from the fitted values: $\left|\hat{R}_{fk_1} - \hat{R}_{fk_2}\right|$. This column highlights that the discrepancies are similarly large between the fitted data and the actual ratings. It is reassuring that similar patterns appear: The two columns are correlated at 0.95, the highest discrepancies are between KLD and the other four rating agencies, and the smallest discrepancies are between Sustainalytics, Vigeo-Eiris, and Asset4. The column "Residuals" shows the mean absolute deviation of the differences between the "True" and the "Fitted" discrepancies: $|\epsilon_{fa} - \epsilon_{fb}|$. The column shows that the errors in the estimation of the aggregation rules are about one third of the variation of the actual scores (71 percent and 24 percent). It is reassuring that these estimation errors do not seem to have a clear pattern that could drive the decomposition results.

Even though the arithmetic decomposition is not a variance decomposition, it is still interesting to document how much of the absolute variation is explained by each type of error. Table 7, last row, indicates that measurement, scope and weight account for 0.52, 0.32, and 0.26 standard deviations. Which corresponds to 47, 29, and 24 percent, respectively. In other words, measurement discrepancy account for a little less half of the discrepancies. In fact, the three sources of divergence are negatively correlated with each other, i.e., the absolute variation of each variable is higher than the absolute variation of the sum of the variables. Even though equation 6 is exact, the equation in absolute values is not.[29] In fact, the mean absolute deviation of scope, measurement, and weight added together are consistently higher than the "Fitted" and "True" mean absolute deviation. These correlations are the reason why this methodology falls short of a proper variance decomposition. The next section proposes a different methodology to cope with this caveat.

5.1.2 Regression Based Decomposition

In this section we present an alternative methodology to decompose the ratings into scope, measurement, and weight divergence. Here we address the shortcoming of the methodology from the previous section, namely the fact that the three sources of divergence are correlated. To do so, we regress the fitted ratings of one agency on the fitted ratings of another and add variables for scope, measurement, and weight by combining information from the two raters.

Lets define the following variables:

Definition 4 ***Measurement, Scope, and Weight Variables***

$$\begin{aligned} Scope_{fa,b} &= C_{fbj_{b,ex}} \cdot \hat{w}_{bj_{b,ex}} && (7) \\ Meas_{fa,b} &= C_{fbj_{com}} \cdot \hat{w}_{aj_{com}} && (8) \\ Weight_{fa,b} &= C_{faj_{com}} \cdot \hat{w}_{bj_{com}} && (9) \end{aligned}$$

The weights in all three terms are calculated based on the reverse-engineering in section 4.2. $Scope_{fa,b}$ is the fitted rating using only the categories and the corresponding weights that are exclusive to rater b. $Meas_{fa,b}$ is the fitted rating using the category scores in rater b for the common category scores and rater a's corresponding weights. Finally, the variable $Weight_{fa,b}$ represents the fitted rating using the common category scores from rater a and the corresponding weights from rater b. Our purpose is to compute the linear regression in Equation 10 and to evaluate the marginal R^2 of the three terms adding them to the regression one at a time.

$$\hat{R}_{fb} = \alpha + \beta \cdot \hat{R}_{fa} + \beta_s \cdot Scope_{fa,b} + \beta_m \cdot Meas_{fa,b} + \beta_w \cdot Weight_{fa,b} + \epsilon \tag{10}$$

The fitted rating $\hat{R}_{fb}$ is the outcome of the the dot product between the category scores C_{fbj} and rater b's estimated weights $\hat{w}_{bj}$; similarly for rating agency a. Let us recall that the fitted rating of rater a is $\hat{R}_{fa} = C_{faj_{com}} \cdot \hat{w}_{aj_{com}} + C_{faj_{ex}} \cdot \hat{w}_{aj_{ex}}$. It follows that $\hat{R}_{fa}$ can be thought of as a control variable for the information that comes from rater a in the construction of the three variables $Scope_{fa,b}$, $Meas_{fa,b}$ and $Weight_{fa,b}$. Hence, $Meas_{fa,b}$ can be attributed to measurement as we already control for the common categories and weights from rater a but not for the common categories from rater b. The same idea is behind $Weight_{fa,b}$ where we already control for the common categories and weights of rater a but not for the weights from rater b. This variable can thus be attributed to weight.

Given the fact that the three terms scope, measurement, and weight are correlated with each other, the order we add them as regressors to Regression 10 matters. We thus run partialing-out regressions in order to calculate a lower and an upper bound of the additional explanatory power

[29] In other words, $\left|\hat{R}_{fa} - \hat{R}_{fb}\right| \neq |\Delta_{scope}| + |\Delta_{meas}| + |\Delta_{weight}|$.

of those terms. For example, to estimate the contribution of scope, we run different comparisons. We estimate two regressions, one with and another without *Scope* to compute the difference in the R^2's. By changing the regressors in the baseline, the contribution of scope changes. We compute the maximum and the minimum of those contributions. In particular, for scope we estimate the following 8 regressions:

$$
\begin{aligned}
\hat{R}_{fb} &= \alpha + \beta \cdot \hat{R}_{fa} && + \epsilon_0 \implies R_0^2 \\
\hat{R}_{fb} &= \alpha + \beta \cdot \hat{R}_{fa} + \beta_s \cdot Scope_{fa,b} && + \epsilon_1 \implies R_1^2 \\
\hat{R}_{fb} &= \alpha + \beta \cdot \hat{R}_{fa} + \beta_m \cdot Meas_{fa,b} && + \epsilon_2 \implies R_2^2 \\
\hat{R}_{fb} &= \alpha + \beta \cdot \hat{R}_{fa} + \beta_s \cdot Scope_{fa,b} + \beta_m \cdot Meas_{fa,b} && + \epsilon_3 \implies R_3^2 \\
\hat{R}_{fb} &= \alpha + \beta \cdot \hat{R}_{fa} + \beta_w \cdot Weight_{fa,b} && + \epsilon_4 \implies R_4^2 \\
\hat{R}_{fb} &= \alpha + \beta \cdot \hat{R}_{fa} + \beta_s \cdot Scope_{fa,b} + \beta_w \cdot Weight_{fa,b} && + \epsilon_5 \implies R_5^2 \\
\hat{R}_{fb} &= \alpha + \beta \cdot \hat{R}_{fa} + \beta_m \cdot Meas_{fa,b} + \beta_w \cdot Weight_{fa,b} && + \epsilon_6 \implies R_6^2 \\
\hat{R}_{fb} &= \alpha + \beta \cdot \hat{R}_{fa} + \beta_s \cdot Scope_{fa,b} + \beta_m \cdot Meas_{fa,b} + \beta_w \cdot Weight_{fa,b} && + \epsilon_7 \implies R_7^2
\end{aligned}
$$

The contribution of scope is the four differences $\{R_1^2 - R_0^2, R_3^2 - R_2^2, R_5^2 - R_4^2, R_7^2 - R_6^2\}$. These differences represent the additional contribution in explanatory power when scope is included.

We present the results in Table 8. For instance, the first row "KLD on Vigeo-Eiris" is the decomposition explaining the KLD rating using Vigeo-Eiris information. The first column presents the baseline R^2. This is simply regressing the KLD rating on the Vigeo-Eiris rating. The first column should be related to the correlation in the fitted ESG ratings across rating agencies. The average fit is 0.40 and fluctuates between 0.16 and 0.57. Notice that the KLD rating is the worst-explained rating. Other ratings explain at most 0.27 of the KLD variation.

Table 8. Range of Variance Explained

			Measurement		Scope		Weight	
	Baseline	**All Covariates**	**Min**	**Max**	**Min**	**Max**	**Min**	**Max**
KLD on Vigeo	22.7%	77.4%	18.0%	22.5%	29.8%	36.6%	0.1%	2.9%
KLD on Sustanalytics	20.9%	77.0%	31.9%	38.2%	17.9%	24.2%	0.0%	1.1%
KLD on RobecoSAM	26.8%	65.6%	22.2%	22.9%	15.6%	16.2%	0.0%	0.4%
KLD on Asset4	19.5%	72.2%	43.1%	48.3%	2.5%	4.8%	1.4%	5.8%
Vigeo on KLD	16.1%	96.4%	12.7%	60.3%	13.2%	60.9%	0.0%	0.3%
Vigeo on Sustainalytics	47.2%	96.0%	8.9%	32.7%	16.1%	39.9%	0.0%	0.0%
Vigeo on RobecoSAM	55.4%	89.2%	21.6%	25.0%	6.6%	11.6%	0.4%	2.3%
Vigeo on Asset4	56.9%	91.9%	29.8%	34.9%	0.1%	4.2%	0.0%	1.8%
Sustainalytics on RobecoSAM	26.8%	89.7%	8.3%	35.2%	26.9%	54.1%	0.5%	0.8%
Sustainalytics on KLD	55.4%	87.1%	3.5%	11.6%	16.8%	27.5%	0.3%	4.2%
Sustainalytics on Vigeo	49.2%	89.1%	5.2%	17.0%	21.1%	34.2%	0.2%	2.4%
Sustainalytics on Asset4	53.1%	77.1%	9.7%	15.3%	1.4%	3.5%	6.8%	11.7%
RobecoSAM on Sustainalytics	19.5%	96.9%	12.0%	61.7%	15.7%	65.4%	0.0%	0.1%
RobecoSAM KLD	56.9%	95.5%	17.2%	31.8%	6.6%	21.2%	0.1%	0.5%
RobecoSAM Vigeo	46.6%	97.1%	13.9%	39.1%	11.4%	36.6%	0.0%	0.6%
RobecoSAM on Asset4	53.1%	86.2%	13.9%	25.8%	6.7%	15.8%	0.5%	4.4%
Asset4 on RobecoSAM	19.5%	96.9%	12.0%	61.7%	15.7%	65.4%	0.0%	0.1%
Asset4 on Sustainalytics	56.9%	95.5%	17.2%	31.8%	6.6%	21.2%	0.1%	0.5%
Asset4 on Vigeo	46.6%	97.1%	13.9%	39.1%	11.4%	36.6%	0.0%	0.6%
Asset4 on KLD	53.1%	86.2%	13.9%	25.8%	6.7%	15.8%	0.5%	4.4%
Average	40.1%	88.0%	16.4%	34.0%	12.4%	29.8%	0.5%	2.2%

This table shows the additional R^2 from regressions of rating a on rating b and scope, measurement, and weight terms from definition 4. We report the maximum and minimum R^2's.

The second column is the R^2 based on a regression that includes all four covariates, i.e., it includes the fitted rating of rater a plus the scope, measurement, and weight variables. The R^2 fluctuates between 0.66 and 0.97 with an average of 0.88. The additional variables in this regression improve the fit by 0.48 on average. The next six columns indicate the minimum and maximum R^2 gain of explanatory power due to including the scope, measurement, and weight variables. The measurement variable is on average the one that contributes the most. In fact, in all of the regressions we estimate, measurement contributes with 0.25 to the increase of the R^2. It fluctuates between 0.16 and 0.34. With 52.9 percent of the total improvement in the fit, more than half of the explanation in the discrepancy is coming from the differences in measurement. The second biggest contribution comes from scope. The average improvement of scope is 0.21, fluctuating from an average minimum of 0.12 to 0.3. On average, scope represents 44.2 percent of the R^2 improvement. Finally, weight is the smallest contributor. It explains on average 0.1 of the increase in R^2, fluctuating between 0 and 0.02, with a share of 2.9 percent of the overall improvement.

This variance decomposition between measurement, scope and weight of 53, 44, and 3 percent respectively is similar to the absolute variation computed in the previous subsection. The results are very similar for the two different decomposition approaches. Measurement is in both the predominant source of divergence, followed by scope and weight, respectively. Even on a more detailed level, both methodologies give similar results.

5.2 Rater Effect

In this section we explore the presence of a *Rater Effect*[30]. The process of evaluating firms' ESG attributes, e.g. human rights, community and society, labor practices, etc., involves judgement calls by the rating agencies. The rater effect implies that those judgements will be correlated with each other. In other words, when the judgement of a company is positive for one particular indicator, it is also likely to be positive for another indicator. One explanation is that rating agencies are mostly organized by firms rather than indicators. A firm that is perceived as good will be seen through a positive lens and receive better indicator scores than the individual indicator would have allowed for, and vice versa. While speaking to RobecoSam we learned about another potential cause for such a rater effect. Some raters make it impossible for firms to receive a good indicator score if they do not give an answer to the corresponding question in the questionnaire. This happens regardless of the actual indicator performance. The extent to which the firms answer specific questions is very likely correlated across indicators. Hence, the willingness to disclose might also explain parts of the rater effect. Technically, the rater effect implies that the discrepancies across categories within a rater are positively correlated. We evaluate the rater effect using two procedures. First, we estimate fixed effects regressions comparing categories, firms and raters. Second, we estimate within rating agency contribution for each of its categories.

5.2.1 Rater Fixed Effects

The first procedure is based on a simple fixed effects decomposition. A firm's score in a given category depends on the firm itself, on the rating agency, and on the category being analyzed. We examine to which extent those three sources explain the variability of scores. We perform the following set of fixed effects regressions:

$$C_{fkj} = \alpha_f \mathbb{1}_f + \epsilon_{fkj,1} \tag{11}$$

$$C_{fkj} = \alpha_f \mathbb{1}_f + \gamma_{fk} \mathbb{1}_{f\times k} + \epsilon_{fkj,2} \tag{12}$$

$$C_{fkj} = \alpha_f \mathbb{1}_f + \gamma_{fj} \mathbb{1}_{f\times j} + \epsilon_{fkj,3} \tag{13}$$

$$C_{fkj} = \alpha_f \mathbb{1}_f + \gamma_{fk} \mathbb{1}_{f\times k} + \gamma_{fj} \mathbb{1}_{f\times j} + \epsilon_{fkj,4} \tag{14}$$

[30] See Shrout and Fleiss (1979), Mount et al. (1997), Griffin and Tang (2011), Griffin et al. (2013) and Fong et al. (2014) for different examples in the literature where rater effects have been evaluated.

where $\mathbb{1}_f$ are dummies for each firm, $\mathbb{1}_{f\times k}$ is an interaction term between firm and rater fixed effects, and $\mathbb{1}_{f\times j}$ is an interaction term between firm and category fixed effects. C_{fkj} is a vector that stacks all cross-sectional scores for all common categories across raters. We drop pure category and rater fixed effects because of the normalization at the rating and category scores level. We only use the intersection of categories from all raters and the common sample of firms to reduce sample bias. We obtain very similar results by including all categories from all raters.

We compute the contribution of the different fixed effects. The baseline regression (eq 11) explains category scores with firm dummies. The second regression adds the assesment that each rater has at the firm level, namely the rater × firm fixed effects. The increment is the rater effect. The third regression uses the firm fixed effects and the category × firm fixed effects. The difference between 13 and 11 is the explanatory power that categories have on the overall firm rating. Finally, equation 14 adds rater × firm fixed effects to equation 13. If the rater effect is zero, the difference in R^2 between the two first regressions and the last two regressions should also be zero. The results of these regressions are shown in Table 9.

Table 9. Investigation of Category and Rater Effect.

Dummies	R^2
Firm	18%
Firm + Firm-Rater	34%
Firm + Firm-Category	44%
Firm + Firm-Category + Firm-Rater	58%

The dependent variable is a vector that stacks all the common category scores for all raters using the common sample.

Two main results emerge. First, firm fixed effects explain 18 percent of the scores. When the variables for the assessment of the rater effect are included it almost doubles to 34 percent (16 percent increase). Similarly, the difference in R^2 between equation 13 and equation 14 yields an increase of 14 percent. Therefore, the rater effect explains about 14 to 16 percent, while the firm fixed effects account for 18 percent. Second, the categories matter. Comparing the estimates of equation 13 versus 11, we find that including categories improves the fit by 26 percent. An alternative way to compute the contribution of the category effect is to compare the outcomes of regressions 14 and 12. The result is a similar increase of the R^2 by 24 percent. Notice that in this simple setting, slightly less than 60 percent of the category scores can be explained with dummies. Even though the rater effect is smaller than the other two, it is clear that it is not irrelevant nor inconsequential. In other words, the rater effect is of the same order of magnitude as the idiosyncratic characteristics of the firms.

5.2.2 LASSO Approach to Rater Effect

We explore the rater effect using an alternative procedure. Here, we concentrate exclusively on the within-rater variation. A rating agency with no rater effect is one in which the correlations between categories are relatively small, a rating agency with strong rater effect implies that the correlations are high. These correlations, however, cannot be accurately summarized by pairwise comparisons. Instead, we can test for the correlations across categories using LASSO regressions. The idea is that a strong rater effect implies that the marginal explanatory power of each category within a rater is diminishing when added one after another. This implies that one could replicate an overall rating with less than the full set of categories.

We test this by re-estimating the linear aggregation rules adding a LASSO penalty. The LASSO regression adds a regularization to the minimization problem of ordinary least squares. The objective is to reduce the number of $w_{kj} \neq 0$ and find the best combination of regressors that maximize the explanatory power of the regression. The optimization is as follows:

$$\min_{w_{kj}} \sum_j (R_{fk} - C_{fkj} * w_{kj})^2 + \lambda \cdot \sum_j |w_{kj}|.$$

where λ controls the penalty. When $\lambda = 0$ the estimates from OLS are recovered. As λ increases, the variables with the smallest explanatory power are eliminated. In other words, the first category that has the smallest explanatory R^2 is dropped from the regression (or its coefficient is set to zero). When λ continues to increase, more and more coefficients are set to zero, until there is only one category left. The simplicity of the LASSO estimation is that instead of running hundreds of regressions and sorting them, the optimization already finds the best combination.

Table 10. Lasso Regressions

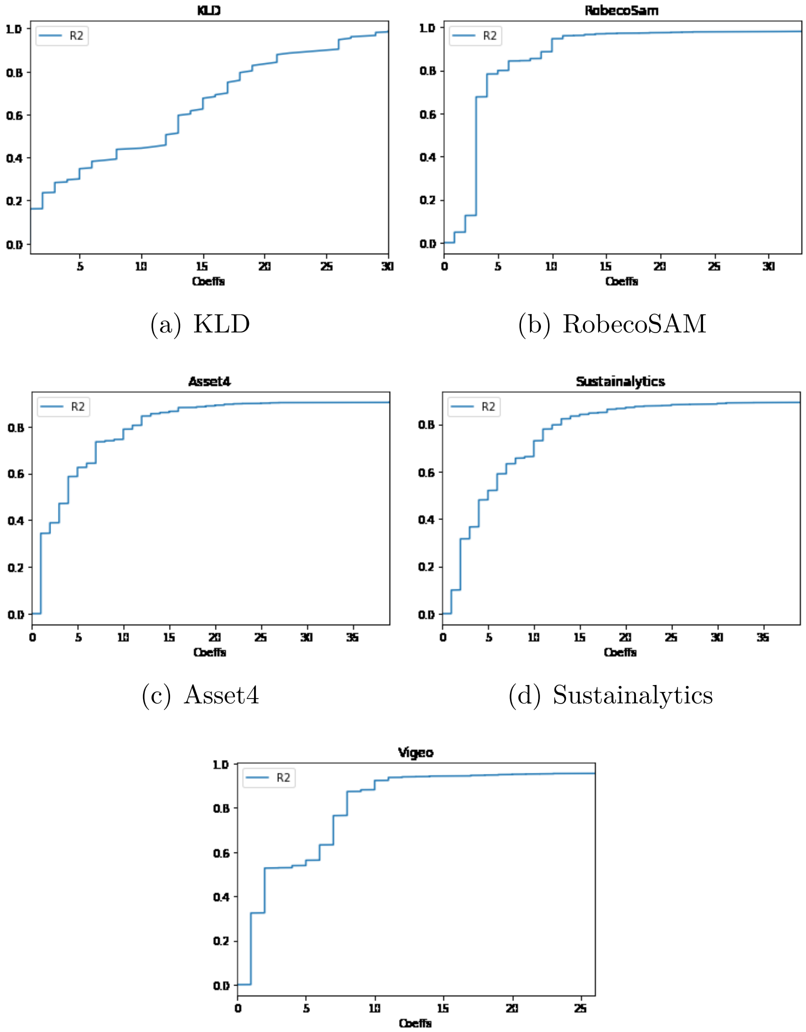
Categories Included	**Vigeo-Eiris**	**RobecoSAM**	**Asset4**	**KLD**	**Sustainalytics**
1	0.31	0.13	0.09	0.11	0.15
2	0.42	0.28	0.28	0.23	0.21
3	0.48	0.77	0.37	0.28	0.28
4	0.65	0.77	0.42	0.30	0.37
5	0.68	0.77	0.55	0.30	0.43
6	0.71	0.84	0.57	0.34	0.53
7	0.76	0.84	0.57	0.34	0.56
8	0.83	0.88	0.58	0.36	0.60
9	0.91	0.96	0.59	0.39	0.63
10	0.93	0.96	0.61	0.44	0.64
15	0.96	0.97	0.81	0.68	0.81
20	0.96	0.98	0.84	0.86	0.83

This table shows the R^2 of a series of lasso regressions of aggregate rating (ESG) of the different rating agencies on the categories of the same rater. The column is the number of indicators that are used as covariates to obtain the corresponding R^2. The highlighted cells represent the number of indicators that constitute 10 percent of the indicators of the particular rating agency.

The objective is to evaluate how much each category contributes to the overall explanatory power. Table 10 shows the rating agencies in the columns and the number of regressors in the rows. For example, the first row documents the R^2 of the category that maximizes the R^2 for a given rater. The second row indicates the R^2 when two categories are included. As expected, the R^2 increases. We proceed until all the categories are included in the regression. The larger the rater effect is, the steeper is the increase in the R^2 explained by the first categories. This is because the initial categories incorporate the rater effect, while the last categories only contribute to the R^2 by their orthogonal component.

In the computation of the aggregation rules (Table 6), the number of categories including the unclassified indicators covered by Vigeo-Eiris, RobeccoSAM, Asset4, KLD, and Sustainalytics are 28, 44, 92, 42, and 64, respectively. 10 percent of the possible regressors therefore are 3,4,9,4, and 6, respectively. We have highlighted these fields in Table 10. Hence, 10 percent of the categories explain 48 percent of the variation in Vigeo-Eiris's ratings, 77 percent in RobeccoSAM, 59 percent in Asset4, only 30 percent in KLD, and 53 percent in Sustainalytics. This illustrates the presence of a rater effect.

For completeness, in Figure 7, we present the increase in the R^2 for each rating agency for all possible categories. The curves reflect the evolution of the R^2. The last part of the curve to the



(a) KLD

(b) RobecoSAM

(c) Asset4

(d) Sustainalytics

(e) Vigeo-Eiris

Figure 7. R^2 of a series of lasso regressions of the aggregate rating (ESG) of the different rating agencies on the categories of the same rater. The x-axis shows how many indicators are used as covariates and the y-axis how much of the variance of the ESG rating they explain.

right coincides with the unrestricted OLS estimates when all variables are included[31]. These figures provide the same message we obtained from the simple statistic of observing the R^2 before. KLD has the smallest cross-category correlation, judging by the slope in Figure 7(a). Sustainalitycs is the second steepest, followed by Vigeo-Eiris and Asset 4, and leaving RobecoSAM as the rating agency where the smallest proportion of categories explains the highest proportion of the ESG rating.

6 Conclusions

The key contribution of this article is to quantitatively disentangle the different drivers of divergence between ESG ratings. The analysis shows that on average, differences in measurement explain 53 percent of the total differences between ESG ratings. Differences in weight explain 3 percent, and differences in scope explain on average 44 percent. Hence, raters disagree both on the extent of the definition of ESG, as much as they disagree on how the various aspects of ESG are measured. We also document a rater effect. The process of evaluating firms' ESG attributes, e.g. human rights, community and society, labor practices, etc., involves judgment calls by the rating agencies. The presence of the rater effect implies that those judgments are correlated with each other. In other words, when the judgment of a company is positive for one particular indicator, it is also likely to be positive for any other indicator of the same rater, and vice versa.

Our methodology allows companies to understand why they received different ratings from different rating agencies. For example, using the results from our arithmetic decomposition, the Korean electronics manufacturer Samsung Electronics Ltd. received a (normalized) rating of 0.94 from Asset4 and -2.32 from KLD, i.e., a difference of 3.52 which is substantial given that both ratings were standardized to have a variance of 1. This difference is composed of 1.90 due to measurement divergence, 1.34 due to weight divergence, and 0.28 due to scope divergence. Further investigation reveals that 0.77 of the measurement divergence is due to a lower assessment by KLD in the category health and safety, and 0.57 due to a lower assessment in the category environmental management system. Of the weight divergence, 0.52 are due to KLD putting greater weight on the category supply chain, and 0.37 from greater weight on the category child labor. In other words, more than two-thirds of the rating divergence is explained by a small number of factors. The categorization in the taxonomy and the simple linearity in the approximation of the aggregation rules allows us to provide a clearer depiction of the reasons behind the rating divergence.

The results have important implications for research, investors, companies, and rating agencies. Researchers should carefully choose the data that underlies future studies involving ESG performance. Some of the results that have been obtained on the basis of a data set might not be replicable with the ratings of another rating agency. In particular, the results indicate that the divergence is most pronounced for KLD data, on which the majority of academic research is based. The robustness of results with regards to the choice of ESG rating is an important step for future research. Some recent studies have included alternative ratings as a robustness check in their empirical analysis, e.g. (Liang and Renneboog, 2016). While this is a reasonable measure, it implicitly assumes that the rating divergence is simply noise. We show that this divergence is not merely noise. Since half of the divergence in ratings is coming from aggregation rules, instead of using aggregate data as it is provided, researchers may consider construct their own measures. The taxonomy provided in this article offers a useful starting point.

[31] See Table A.2

For investors, this paper also shows a way to interpret the discrepancy between different ESG ratings by tracing them back to specific differences in scope, measurement, and weight. For instance, investors could reduce the discrepancy between raters by about 50 percent when they impose their own weighting on the indicators of different rating agencies. Remaining differences can be traced to the indicators that are driving the discrepancy, potentially guiding an investor's additional research. This paper introduces a framework under which investors can integrate various ESG ratings into a coherent decision-making process. Nevertheless, until there are more standardized and easily accessible indicators available, investors will be exposed to diverging ESG ratings.

For companies, the results highlight that there is substantial disagreement about their ESG performance. The divergence happens not only at the aggregate level but also in relatively specific sub-categories of ESG performance, such as human rights or energy. This situation might frustrate attempts by companies to improve, because the chance that their efforts are recognized consistently by ESG rating providers is small. In many cases, improving scores with one rating provider is unlikely to result in improved scores at another. Thus, in their current form, ESG ratings do not play a role as important as potentially possible in guiding companies towards improvement. To change the situation, companies should work with rating agencies to establish open and transparent disclosure standards and ensure that the data is publicly accessible. If companies fail to do so, the demand for ESG information will push rating agencies to base the creation of the data on other sources prone to divergence.

Finally, for rating agencies, the paper diagnoses a fundamental problem of the ESG rating industry itself, namely that differences between raters are not merely differences in opinion, but differences in measurement. The presence of the rater effect has implications for the organizational structure of rating agencies. The data shows that one rater's view of a particular company strongly correlates across different categories. Future research should explore why this occurs. Lastly, we find that ESG ratings can be replicated with a dramatically reduced set of indicators. This result may be driven by the rater effect, but it may also point to potential redundancies.

7 References

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

D. Bongaerts, K. J. M. Cremers, and W. N. Goetzmann. Tiebreaker: Certification and Multiple Credit Ratings. *The Journal of Finance*, 67(1):113–152, 2012.

L. Bouten, C. H. Cho, G. Michelon, and R. W. Roberts. CSR Performance Proxies in Large-Sample Studies: 'Umbrella Advocates', Construct Clarity and the 'Validity Police'. *SSRN Electronic Journal*, 2017.

R. Cantor and F. Packer. Differences of opinion and selection bias in the credit rating industry. *Journal of Banking and Finance*, 21(10):1395–1417, 1997.

A. K. Chatterji, D. I. Levine, and M. W. Toffel. How Well Do Social Ratings Actually Measure Corporate Social Responsibility? *Journal of Economics & Management Strategy*, 18(1):125–169, 2009.

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

M. Delmas and V. D. Blass. Measuring Corporate Environmental Performance: the Trade-Offs of Sustainability Ratings. *Business Strategy and the Environment*, 19(4):245–260, 2010.

T. Didier, R. Rigobon, and S. Schmukler. Unexploited gains from international diversification: Patterns of portfolio holdings around the world. 2012.

G. Dorfleitner, G. Halbritter, and M. Nguyen. Measuring the level and risk of corporate responsibility – An empirical comparison of different ESG rating approaches. *Journal of Asset Management*, 16 (7):450–466, 2015.

R. G. Eccles and J. C. Stroehle. Exploring Social Origins in the Construction of ESG Measures. (ID 3212685), 2018.

E. F. Fama and K. R. French. Disagreement, tastes, and asset prices. *Journal of Financial Economics*, 83(3):667–689, 2007.

K. Y. Fong, H. G. Hong, M. T. Kacperczyk, and J. D. Kubik. Do security analysts discipline credit rating agencies? *AFA 2013 San Diego Meeting Paper*, 2014.

R. Gibson and P. Krueger. The Sustainability Footprint of Institutional Investors. SSRN Scholarly Paper ID 2918926, Social Science Research Network, Rochester, NY, 2018.

J. M. Griffin and D. Y. Tang. Did credit rating agencies make unbiased assumptions on cdos? *American Economic Review*, 101(3):125–130, 2011.

J. M. Griffin, J. Nickerson, and D. Y. Tang. Rating shopping or catering? An examination of the response to competitive pressure for cdo credit ratings. *Review of Financial Studies*, 26(9): 2270–2310, 2013.

GSIA. Global Sustainable Investment Review. Technical report, 2018.

L. Güntay and D. Hackbarth. Corporate bond credit spreads and forecast dispersion. *Journal of Banking & Finance*, 34(10):2328–2345, 2010.

J. Jewell and M. Livingston. Split ratings, bond yields, and underwriter spreads. *Journal of Financial Research*, 21(2):185–204, 1998.

P. Krueger, Z. Sautner, and L. T. Starks. The importance of climate risks for institutional investors. 2018.

H. Liang and L. Renneboog. On the Foundations of Corporate Social Responsibility. *The Journal of Finance*, pages 1–59, 2016.

K. V. Lins, H. Servaes, and A. M. Tamayo. Social Capital, Trust, and Firm Performance: The Value of Corporate Social Responsibility during the Financial Crisis. *Journal of Finance*, 2017.

M. K. Mount, M. R. Sytsma, J. F. Hazucha, and K. E. Holt. Rater-ratee rate effects in developmental performance ratings of managers. *Personnel Psychology*, 50(1):51–69, 1997.

N. Semenova and L. G. Hassel. On the Validity of Environmental Performance Metrics. *Journal of Business Ethics*, 132(2):249–258, 2015.

P. E. Shrout and J. L. Fleiss. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428, 1979.

Appendices

A Appendix

Table A.1. Number of observations per criterion.

	KLD	**Vigeo-Eiris**	**RobecoSAM**	**Sustainalytics**	**Asset4**
Access to Basic Services	98	0	0	515	4024
Access to Healthcare	247	0	111	244	122
Animal Welfare	0	0	0	641	290
Anti-Competitive Practices	4295	1418	0	0	4025
Audit	0	2319	0	2450	4025
Biodiversity	4295	713	249	18	4025
Board	0	2319	0	4551	4025
Board Diversity	0	0	0	0	0
Board Gender Diversity	0	0	0	2450	0
Business Ethics	4295	0	1668	4551	4025
Chairman Ceo Separation	0	0	0	4551	4025
Child Labor	4295	5	0	0	4025
Climate Risk Mgmt.	2957	0	1668	0	4024
Clinical Trials	0	0	0	128	220
Collective Bargaining	0	2254	0	2993	4025
Community and Society	2414	1752	1668	4551	4025
Corruption	1396	2077	0	4551	4025
Customer Relationship	4295	839	1310	4551	4025
Discrimination and Diversity	4295	2312	0	4550	4025
ESG incentives	0	0	0	2450	0
Electromagnetic Fields	0	0	49	64	0
Employee Development	1592	2102	1668	91	4025
Employee Turnover	0	0	0	2448	1171
Energy	116	2213	136	2531	4024
Environmental Fines	0	0	0	2450	4025
Environmental Mgmt. System	2032	0	0	4551	692
Environmental Policy	0	2319	1668	4551	4025
Environmental Reporting	0	0	1668	3785	4024
Financial Inclusion	467	0	0	776	0
Forests	0	0	8	33	0
GHG Emissions	4295	823	0	2466	4024
GHG Policies	0	0	41	4551	4024
GMOs	0	0	105	249	305
Global Compact Membership	0	0	0	4550	4025
Green Buildings	338	0	114	447	4024
Green Products	1198	677	410	2837	4024
HIV Programmes	0	0	0	61	4024
Hazardous Waste	0	0	39	1502	688
Health and Safety	4295	2317	1525	4551	4025
Human Rights	4295	1274	0	4551	4025
Indigenous Rights	495	0	0	494	4025
Labor Practices	4295	2319	1668	2448	4025
Lobbying	0	1470	0	4551	0
Non-GHG Air emissions	0	0	0	1379	3040
Ozone Depleting Gases	0	0	0	122	4024
Packaging	80	0	182	0	0
Philantrophy	0	437	1668	2450	4024
Privacy and IT	530	0	152	380	0
Product Safety	4295	1835	106	4551	4025
Public Health	322	0	205	91	0
Remuneration	4295	2319	62	2450	4025
Reporting Quality	0	0	0	4551	4025
Resource Efficiency	0	0	1666	100	4024
Responsible Marketing	4295	934	181	502	2079
Shareholders	0	2186	0	0	4025
Site Closure	0	0	49	163	0
Supply Chain	4295	1934	1239	4551	4024
Sustainable Finance	4295	0	269	1008	861
Systemic Risk	459	0	164	0	0
Taxes	0	0	1152	2700	3497
Toxic Spills	4295	0	0	241	3041
Unions	2734	0	0	0	1266
Waste	4295	780	49	48	4024
Water	4295	756	275	1895	4024
Sum	94785	42703	23192	125465	174232

Number of observations for each criterion in our taxonomy. We calculate a value for each criterion on the firm level by taking the average of the available indicators for firm f and rater k. As indicators depend on industries the values of the same criterion but for different firms might not use the same indicators as input.

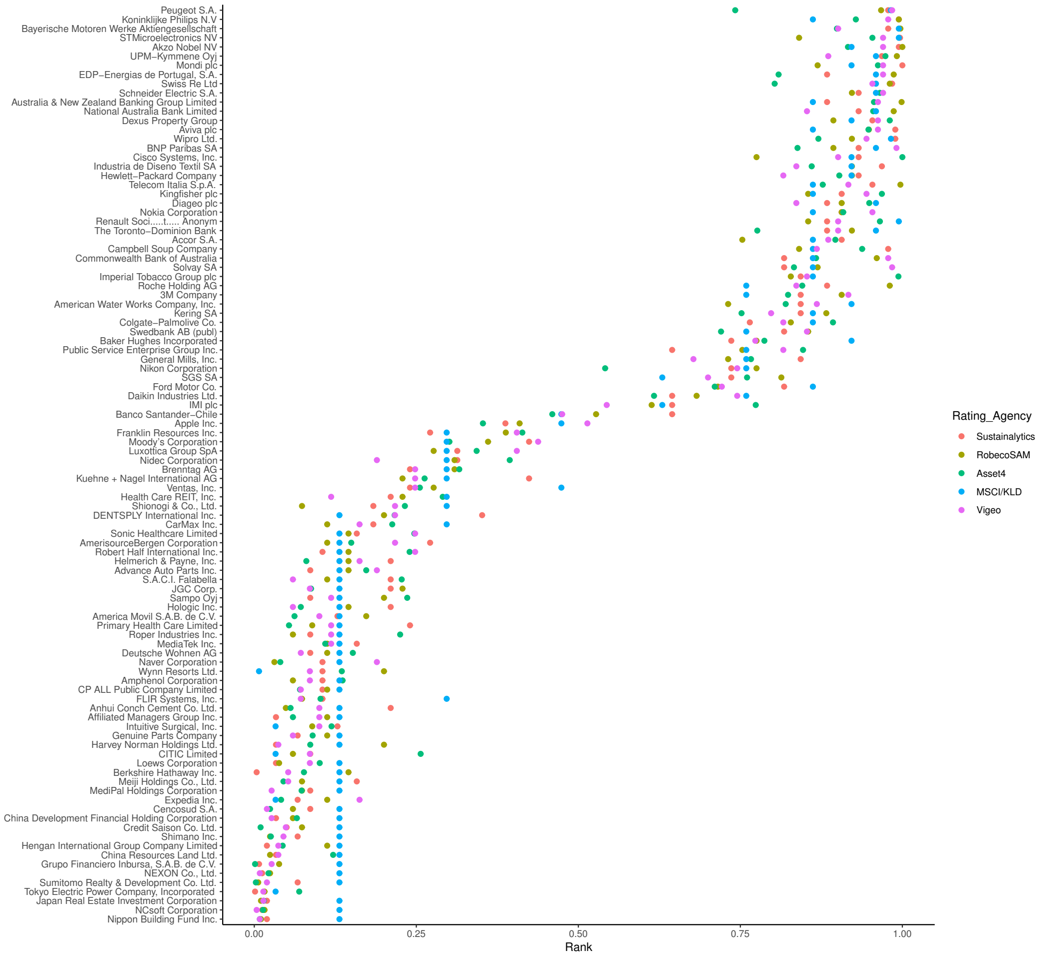


Figure A.1. Comparison of firms' rankings for different rating agencies.

100 firms with the lowest median average distance within the common sample (n=823). Firms within these group have been sorted by their respective median. Each rating agency ranking is plotted in a different color.

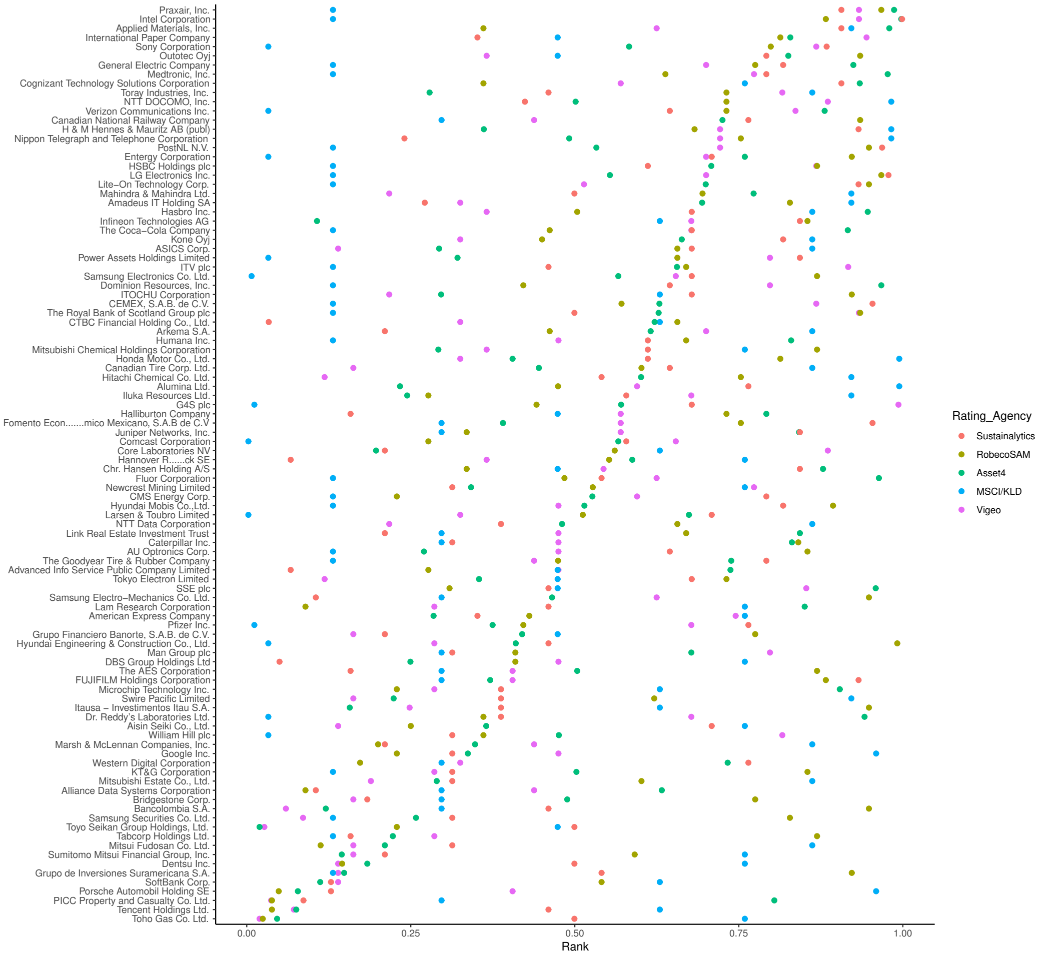


Figure A.2. Comparison of firms' rankings for different rating agencies.

100 firms with the highest median average distance within the common sample (n=823). Firms within these group have been sorted by their respective median. Each rating agency ranking is plotted in a different color.

Table A.2. Aggregation Rule Estimation without parameter restrictions and common sample.

	KLD	Vigeo-Eiris	RobecoSAM	Sustainalytics	Asset4
Access to Basic Services	1.09***	-	-	0.01	-0.02**
Access to Healthcare	0.86***	-	0.01	0.06***	-0.06**
Animal Welfare	-	-	-	0.02***	-0.01
Anti-Competitive Practices	0.99***	0.01*	-	-	-0.11***
Audit	-	0.06***	-	0.00	0.01
Biodiversity	1.71***	0.02***	-0.01	-0.02	0.00
Board	-	0.05***	-	0.05***	0.33***
Board Diversity	0.00***	-	-	0.00***	0.00***
Board Gender Diversity	0.00***	-	-	0.00**	-
Business Ethics	1.02***	-	0.07***	0.06***	-0.03
Chairman Ceo Separation	-	-	-	0.02***	-0.01
Child Labor	0.94***	-0.09	-	-	-0.07
Climate Risk Mgmt.	1.14***	-	0.12***	-	0.04***
Clinical Trials	-	-	-	-0.02***	0.07***
Collective Bargaining	-	0.04***	-	0.02***	-0.10*
Community and Society	1.06***	0.02***	0.06***	0.04***	0.17***
Corruption	1.01***	0.06***	-	0.01***	-0.04***
Customer Relationship	1.92***	0.02***	0.06***	0.07***	0.28***
Discrimination and Diversity	0.81***	0.09***	-	0.04***	0.16***
ESG incentives	-	-	0.00***	0.00	-
Electromagnetic Fields	-	-	-0.05***	0.02***	-
Employee Development	1.07***	0.03***	0.21***	0.02	0.29***
Employee Turnover	-	-	-	0.00**	-0.01**
Energy	1.00***	0.06***	0.04***	0.01***	0.00
Environmental Fines	0.00***	-	-	0.00	-0.09
Environmental Mgmt. System	1.01***	-	-	0.06***	-0.01
Environmental Policy	-	0.09***	0.07***	0.03***	0.07***
Environmental Reporting	-	-	0.03***	0.01***	0.02***
Financial Inclusion	0.94***	-	-	-0.02***	-
Forests	-	-	0.06***	-0.01	-
GHG Emissions	0.99***	0.02***	-	0.00	-0.04***
GHG Policies	-	-	0.01	0.03***	0.12***
GMOs	-	-	0.00	0.01	0.01
Global Compact Membership	-	-	-	0.01***	-0.01
Green Buildings	0.95***	-	0.09***	0.04***	0.02***
Green Products	1.01***	0.05***	0.03***	0.05***	0.22***
HIV Programmes	-	-	-	-0.04***	-0.02**
Hazardous Waste	-	-	-0.05***	0.01*	-0.01
Health and Safety	1.85***	0.08***	0.03***	0.03***	0.03**
Human Rights	4.79***	0.00	0.00***	0.02***	0.13***
Indigenous Rights	0.96***	-	-	0.02***	-0.02
Labor Practices	2.37***	0.12***	0.08***	0.00	0.09
Lobbying	-	-0.02***	0.00	0.04***	-
Non-GHG Air emissions	-	-	-	0.01*	0.02**
Ozone Depleting Gases	-	-	-	0.00	-0.02***
Packaging	0.96***	-	-0.01**	-	-
Philantrophy	0.00	0.05***	0.06***	0.01***	0.01*
Privacy and IT	1.11***	-	0.06***	0.02***	-
Product Safety	2.36***	0.05***	-0.01	0.04***	0.08***
Public Health	1.16***	-	0.01	0.00	-
Remuneration	2.79***	0.06***	0.11***	0.00	0.23***
Reporting Quality	0.00	-	-	0.04***	0.10***
Resource Efficiency	0.00	0.00	0.07***	0.03**	0.09***
Responsible Marketing	1.02***	0.00	0.05***	0.00	0.00
Shareholders	-	0.03***	-	-	0.37***
Site Closure	0.00	0.00	-0.01	0.02	0.00***
Supply Chain	4.28***	0.03***	0.05***	0.13***	0.06***
Sustainable Finance	2.39***	-	0.09***	0.08***	0.06***
Systemic Risk	0.97***	-	0.04***	-	-
Taxes	-	-	0.01	0.02***	0.06***
Toxic Spills	0.98***	-	-	-0.01	-0.01
Unions	1.01***	0.00	0.00	0.00	-0.01
Waste	2.12***	0.02***	0.02	-0.01	0.03***
Water	2.08***	-0.02***	0.03***	0.01***	0.01
Intercept	0.04***	3.15***	-1.58***	11.00***	-99.90***
Unclassified Indicators	Yes	Yes	Yes	Yes	Yes
R2	0.98	0.96	0.98	0.89	0.92
Observations	2714	2319	1668	4551	4025

This table shows the coefficients of ordinary leased squares regressions of aggregate rating (ESG) of a rater k on the categories of the same rater. We use our Taxonomy. We calculate a value for each criterion on the firm level by taking the average of the available indicators for firm f and rater k. As categories depend on industries we fill missing values of the dependent variables with zeros. ***,** and * denote statistical significance at the one, five and ten percent level, respectively. Non-existent category scores are denoted as blanks, whereas redundant category scores with a coefficient very close to zero are denoted as dashes.

Table A.3. Number of indicators per Categories (SASB taxonomy).

	KLD	**Vigeo-Eiris**	**RobecoSAM**	**Sustainalytics**	**Asset4**
GHG Emissions	1	1	2	8	9
Air Quality	0	0	0	2	3
Energy Management	1	1	6	3	5
Water & Wastewater Management	2	1	2	2	3
Waste & Hazardous Materials Management	3	1	3	4	5
Ecological Impacts	3	3	7	11	9
Human Rights & Community Relations	7	2	7	6	16
Customer Privacy	2	0	3	1	0
Access & Affordability	2	0	3	8	2
Product Quality & Safety	6	3	2	2	13
Customer Welfare	4	1	5	3	7
Selling Practices & Product Labeling	1	1	3	3	1
Labor Practices	5	6	1	6	20
Employee Health & Safety	2	1	1	8	8
Employee Engagement, Diversity & Inclusion	9	2	2	5	22
Product Design & Lifecycle Management	2	1	2	7	20
Supply Chain Management	6	4	3	21	4
Materials Sourcing & Efficiency	0	0	3	1	6
Physical Impacts of Climate ChangeÃČÂĆÃĆÂă	2	0	2	0	1
Business Ethics	2	1	2	7	3
Competitive Behavior	1	1	0	0	2
Management of the Legal & Regulatory Environment	1	0	1	3	2
Critical Incident Risk Management	1	0	0	1	2
Systemic Risk Management	1	0	1	0	0
Unclassfied	14	8	19	51	119
Sum	78	38	80	163	282

We consider a category as covered by the rating agency if at least one firm is rated in that category.

Table A.4. Correlation between rating agencies at the level of categories (SASB taxonomy).

	KL:A4	**KL:RS**	**KL:SA**	**KL:VI**	**RS:A4**	**RS:SA**	**SA:A4**	**VI:A4**	**VI:RS**	**VI:SA**	**Average**
GHG Emissions	-0.12		-0.07	-0.05	0.40	0.44	0.63	0.57	0.71	0.34	0.32
Air Quality							0.42				0.42
Energy Management	0.27	0.31	0.12	0.24	0.22	0.26	0.30	0.45	0.37	0.38	0.29
Water & Wastewater Management	0.23	0.20	0.31	0.32	0.12	0.42	0.40	0.40	0.47	0.47	0.33
Waste & Hazardous Materials Management	0.27	0.36	0.36	0.33	0.17	0.20	0.40	0.37	0.46	0.38	0.33
Ecological Impacts	0.43	0.42	0.49	0.40	0.70	0.71	0.65	0.59	0.70	0.66	0.57
Human Rights & Community Relations	0.17	0.16	-0.26	0.23	0.64	-0.12	0.06	0.52	0.54	-0.01	0.19
Customer Privacy		0.32	0.36			0.27					0.32
Access & Affordability	0.45	0.53	0.58		0.48	0.65	0.48				0.53
Product Quality & Safety	0.02	0.19	0.02	0.05	0.37	-0.10	-0.05	0.25	0.49	-0.09	0.11
Customer Welfare	-0.02	-0.04	0.23	-0.09	0.46	-0.13	-0.13	0.52	0.50	-0.06	0.12
Selling Practices & Product Labeling	0.20	-0.34	-0.47	-0.08	-0.11	0.60	-0.07	0.00	0.43	0.40	0.06
Labor Practices	0.16	0.10	0.20	0.26	0.42	0.46	0.40	0.51	0.57	0.56	0.36
Employee Health & Safety	0.28	0.24	0.04	0.30	0.57	-0.15	-0.16	0.71	0.63	-0.14	0.23
Employee Engagement, Diversity & Inclusion	0.15	0.16	0.13	0.18	0.61	0.40	0.55	0.56	0.51	0.58	0.38
Product Design & Lifecycle Management	0.36	0.32	0.26	0.13	0.54	0.37	0.52	0.35	0.38	0.46	0.37
Supply Chain Management	0.16	0.11	0.17	0.17	0.56	0.53	0.53	0.63	0.64	0.56	0.41
Materials Sourcing & Efficiency					0.59	0.33	0.34				0.42
Physical Impacts of Climate Change	0.44	0.45			0.56						0.48
Business Ethics	0.27	0.02	0.05	0.00	-0.18	0.50	-0.13	-0.17	0.57	0.57	0.15
Competitive Behavior	0.55			-0.04				-0.05			0.15
Management Legal & Regulatory Environment					-0.02	0.09	-0.02				0.02
Critical Incident Risk Management	0.03		-0.21				0.07				-0.04
Systemic Risk Management		0.26									0.26
	0.23	0.21	0.13	0.15	0.37	0.30	0.26	0.39	0.53	0.34	

Correlations between the different categories from different rating agencies. We calculate a value for each criterion on the firm level by taking the average of the available indicators for firm f and rater k. As indicators depend on industries the values of the same criterion but for different firms might not use the same indicators as input. The panel is unbalanced due to differences in scope of different ratings agencies and categories being conditional on industries.

The SASB categories of data security and business model resilience are not displayed in this table, because either none or only one of the rating agencies provides indicators for these categories.

Table A.5. Estimates of Non Negative Least Squares Regression using the SASB taxonomy.

	KLD	**Vigeo-Eiris**	**RobecoSAM**	**Sustainalytics**	**Asset4**
GHG Emissions	0.031***	0.045***	0.012***	0.135***	0.008
Air Quality	-	-	-	0.012	0.000
Energy Management	0.058***	0.108***	0.014**	0.017	0.027*
Water & Wastewater ManagementÃĆÂă	0.181***	0.000	0.005	0.048***	0.031**
Waste & Hazardous Materials Management	0.197***	0.009	0.000	0.042***	0.032**
Ecological Impacts	0.220***	0.170***	0.154***	0.273***	0.003
Human Rights & Community Relations	0.314***	0.028***	0.059***	0.110***	0.085***
Customer Privacy	0.118***	-	0.042***	0.036***	-
Access & Affordability	0.071***	-	0.000	0.027**	0.000
Product Quality & SafetyÃĆÂă	0.233***	0.063***	0.000	0.046***	0.052***
Customer Welfare	0.119***	0.031***	0.113***	0.125***	0.088***
Selling Practices & Product Labeling	0.079***	0.008	0.026***	0.000	0.000
Labor Practices	0.203***	0.184***	0.054***	0.082***	0.074***
Employee Health & Safety	0.181***	0.130***	0.050***	0.032**	0.055***
Employee Engagement, Diversity & Inclusion	0.135***	0.190***	0.221***	0.097***	0.143***
Product Design & Lifecycle Management	0.129***	0.017	0.030***	0.160***	0.109***
Supply Chain Management	0.123***	0.051***	0.055***	0.241***	0.055***
Materials Sourcing & Efficiency	-	-	0.104***	0.004	0.134***
Physical Impacts of Climate ChangeÃĆÂă	0.238***	-	0.138***	-	0.073***
Business Ethics	0.164***	0.088***	0.054***	0.134***	0.013
Competitive Behavior	0.134***	0.016*	-	-	0.050***
Management of the Legal & Regulatory Environment	0.000	-	0.003	0.008	0.008
Critical Incident Risk Management	0.103***	-	-	0.000	0.008
Systemic Risk Management	0.111***	-	0.048***	-	-
Unclassified Indicators	Yes	Yes	Yes	Yes	Yes
R2	0.98	0.96	0.98	0.90	0.92
Observations	823	823	823	823	823

Non negative linear regressions (positivity constraints on the coefficients) of aggregate rating (ESG) of a rater k on the categories of the same rater. As categories depend on industries we fill missing values of the dependent variables with zeros. ***,** and * denote statistical significance at the one, five and ten percent level, respectively. As the data was previously normalized we exclude the constant term. The standard errors are bootstrapped. Non-existent categories are denoted as dashes.

Table A.6. Arithmetic Decomposition using the SASB taxonomy

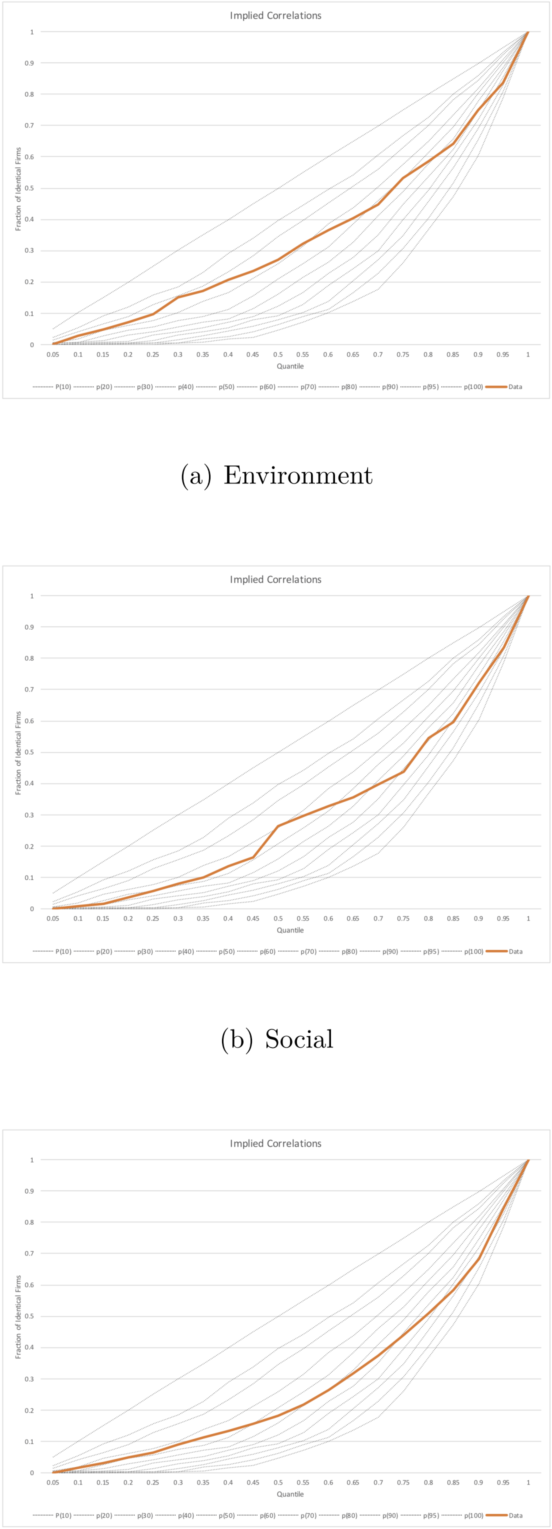
		Scope	**Measurement**	**Weight**	**Residuals**	**Fitted**	**True**
KLD	**Vigeo-Eiris**	0.42	0.66	0.31	0.18	0.80	0.80
KLD	**RobecoSam**	0.28	0.66	0.40	0.15	0.80	0.80
KLD	**Sustainalytics**	0.33	0.64	0.25	0.27	0.73	0.77
KLD	**Asset4**	0.44	0.56	0.48	0.25	0.81	0.87
Vigeo-Eiris	**RobecoSam**	0.34	0.41	0.14	0.17	0.61	0.62
Vigeo-Eiris	**Sustainalytics**	0.28	0.45	0.17	0.29	0.54	0.60
Vigeo-Eiris-Eiris	**Asset4**	0.32	0.39	0.19	0.27	0.55	0.62
RobecoSam	**Sustainalytics**	0.19	0.46	0.24	0.27	0.58	0.65
RobecoSam	**Asset4**	0.33	0.46	0.11	0.25	0.63	0.71
Sustainalytics	**Asset4**	0.35	0.41	0.26	0.33	0.52	0.65
Average		0.33	0.51	0.25	0.24	0.66	0.71

Results from the arithmetic decomposition. First, we estimate the weights by regressing the ESG rating of one rater on the categories of the same rater. Second, we construct two different ratings for rater a and b by only taking categories of mutually exclusive categories and using the weights from step 1. The mean absolute deviation of the difference of those two ratings reflects the differences in the scope between the two rating agencies. Third, we stack the two firm-by-categories matrices of the common categories between rater a and b on each other and calculate a new set of weights that is thus common to both raters. We then subtract the fitted ratings of rater b from rater a in the common categories and calculate the mean absolute deviation to determine the divergence in measurement. Fourth, we calculate the divergence stemming from the aggregation weights by subtracting the residuals from the previous step of rater b from rater a and calculate the mean absolute deviation. The last column reports the mean absolute deviation of the residuals of the estimation procedure.

Table A.7. Range of Variance Explained using the SASB taxonomy

			Measurement		**Scope**		**Weight**	
	Baseline	**All Covariates**	**Min**	**Max**	**Min**	**Max**	**Min**	**Max**
KLD on Vigeo	21.73%	75.84%	27.52%	37.25%	15.05%	21.95%	1.75%	5.19%
KLD on Sustanalytics	20.20%	72.88%	38.75%	44.48%	7.81%	13.94%	0.00%	0.74%
KLD on RobecoSAM	25.36%	79.86%	34.25%	35.34%	18.86%	20.20%	0.01%	0.44%
KLD on Asset4	18.70%	80.17%	41.79%	53.05%	7.27%	14.53%	0.20%	8.79%
Vigeo on KLD	16.10%	95.82%	40.49%	58.65%	15.30%	33.15%	0.05%	0.95%
Vigeo on Sustainalytics	47.90%	96.32%	25.31%	37.66%	10.66%	23.11%	0.00%	0.44%
Vigeo on RobecoSAM	55.33%	96.38%	23.59%	27.39%	12.58%	17.05%	0.03%	2.64%
Vigeo on Asset4	55.83%	94.85%	23.26%	30.00%	8.71%	14.45%	0.19%	6.93%
Sustainalytics on RobecoSAM	25.36%	86.56%	34.07%	48.55%	12.55%	25.77%	0.00%	1.84%
Sustainalytics on KLD	55.33%	87.50%	20.38%	26.53%	4.12%	9.61%	0.80%	4.08%
Sustainalytics on Vigeo	49.72%	88.78%	25.61%	33.39%	4.75%	11.02%	0.52%	3.32%
Sustainalytics on Asset4	53.79%	87.08%	18.05%	23.19%	6.81%	12.32%	1.61%	5.79%
RobecoSAM on Sustainalytics	18.70%	98.35%	21.53%	57.30%	22.30%	57.92%	0.00%	0.21%
RobecoSAM KLD	55.83%	97.83%	12.42%	26.20%	15.35%	29.13%	0.00%	2.96%
RobecoSAM Vigeo	45.68%	97.72%	20.85%	35.11%	16.45%	30.70%	0.14%	1.12%
RobecoSAM on Asset4	53.79%	97.08%	16.85%	21.53%	21.62%	26.30%	0.00%	0.59%
Asset4 on RobecoSAM	18.70%	98.35%	21.53%	57.30%	22.30%	57.92%	0.00%	0.21%
Asset4 on Sustainalytics	55.83%	97.83%	12.42%	26.20%	15.35%	29.13%	0.00%	2.96%
Asset4 on Vigeo	45.68%	97.72%	20.85%	35.11%	16.45%	30.70%	0.14%	1.12%
Asset4 on KLD	53.79%	97.08%	16.85%	21.53%	21.62%	26.30%	0.00%	0.59%
Average	39.67%	91.20%	24.82%	36.79%	13.79%	25.26%	0.27%	2.54%

This table shows the additional R^2 from regressions of rating a on rating b and scope, measurement, and weight terms from definition 4. We report the maximum and minimum R^2's.



(a) Environment

(b) Social

(c) Governance

Figure A.3. Quantile Ranking Counts for E,S,G and ESG for all Raters

The gray lines represent simulated data for each quantile from 10 to 100 percent, i.e., an implicit correlation of 10 to 100 percent. The orange line is the quantile ranking count for the true data. i.e., the fraction of identical companies in the sub sample of a given quantile.