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

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

May 17, 2020

Abstract

This paper investigates the divergence of environmental, social, and governance (ESG) ratings. Based on data from six prominent rating agencies—namely, KLD (MSCI Stats), Sustainalytics, Vigeo Eiris (Moody's), RobecoSAM (S&P Global), Asset4 (Refinitiv), and MSCI—we decompose the divergence into three sources: different scope of categories, different measurement of categories, and different weights of categories. We find that scope and measurement divergence are the main drivers, while weights divergence is less important. In addition, we detect a rater effect where a rater's overall view of a firm influences the assessment of specific categories.

We thank Jason Jay, Kathryn Kaminski, Eric Orts, Robert Eccles, Yannick Le Pen, Andrew King,

Environmental, social, and governance (ESG) rating providers[1] have become influential institutions. Investors with over $80 trillion in combined assets have signed a commitment to integrate ESG information into their investment decisions (PRI, 2018). Many institutional investors expect corporations to manage ESG issues (Krueger et al., 2020) and monitor their holdings' ESG performance (Dyck et al., 2019). Sustainable investing is growing fast and mutual funds that invest according to ESG ratings experience sizable inflows (Hartzmark and Sussman, 2019). Due to these trends, more and more investors rely on ESG ratings to obtain a third-party assessment of corporations' ESG performance. There are also a growing number of academic studies that rely on ESG ratings for their empirical analysis (see, for example, Liang and Renneboog (2017), Servaes (2013), Hong and Kostovetsky (2012), and Lins et al. (2017)). As a result, ESG ratings increasingly influence financial decisions, with potentially far-reaching effects on asset prices and corporate policies.

However, ESG ratings from different providers disagree substantially (Chatterji et al., 2016). In our data set of ESG ratings from six different raters—namely, KLD (MSCI Stats), Sustainalytics, Vigeo Eiris (Moody's), RobecoSAM (S&P Global), Asset4 (Refinitiv), and MSCI—the correlations between the ratings are on average 0.54, and range from 0.38 to 0.71. This means that the information that decision-makers receive from ESG rating agencies is relatively noisy. Three major consequences follow: First, ESG performance is less likely to be reflected in corporate stock and bond prices, as investors face a challenge when trying to identify outperformers and laggards. Investor tastes can influence asset prices (Fama and French, 2007; Hong and Kacperczyk, 2009; Pastor et al., 2020), but only when a large enough fraction of the market holds and implements a uniform nonfinancial 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 hampers 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

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 between existing ESG ratings. To do so, we specify the ratings as consisting of three basic elements: (1) a scope, which denotes all the attributes that together constitute the overall concept of ESG performance; (2) indicators that yield numerical measures of the attributes; and (3) an aggregation rule that combines the indicators into a single rating.

On this basis, we identify three distinct sources of divergence. *Scope divergence* refers to the situation where ratings are based on different sets of attributes. Attributes such as carbon emissions, labor practices, and lobbying activities may, for instance, be included in the scope of a rating. One rating agency may include lobbying acivities, while another might not, causing the two ratings to diverge. *Measurement divergence* refers to a 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-related court cases taken against the firm. Both capture aspects of the attribute labor practices, but they are likely to lead to different assessments. Indicators can focus on policies, 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. Finally, *weights divergence* emerges when rating agencies take different views on the relative importance of attributes. For example, the labor practices indicator may enter the final rating with greater weight than the lobbying indicator. The contributions of scope, measurement, and weights divergence are all intertwined, which makes it difficult to interpret the divergence of aggregate ratings. Our goal is to estimate to what extent each of the three sources drives the overall divergence of ESG ratings.

Methodologically, we approach the problem in three steps. First, we categorize all 709 indicators provided by the different data providers into a common taxonomy of 65 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

that matches indicators by attribute. We created the taxonomy starting from the population of 709 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 unclassified. As such, the taxonomy approximates the population of common attributes as granularly as possible and across all raters. Based on the taxonomy, we calculate rater-specific category scores by averaging indicators that were assigned to the same category. Second, we regress the original rating on those category scores, using a non-negative least squares regression, where coefficients are constrained to be equal to or larger than zero. The regression models yield fitted versions of the original ratings, and we can compare these fitted ratings to each other in terms of scope, measurement, and aggregation rule. Third, we calculate the contribution of divergence in scope, measurement, and weights to 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 of 79 to 99% on the basis of our common taxonomy. This demonstrates that although rating agencies take very different approaches, it is possible to fit them into a consistent framework that reveals in detail how much and for what reason ratings differ. We use linear regressions, neural networks, and random forests to estimate aggregation rules, but it turns out that a simple linear regression is in almost all cases the most efficient method.

Second, we find that measurement divergence is the main driver of rating divergence, closely followed by scope divergence, while weights divergence plays a minor role. This means that users of ESG ratings, for instance financial institutions, cannot easily resolve discrepancies between two raters by readjusting the weights of individual indicators. Instead, rating users have to deal with the problem that the divergence is driven both by what is measured and by how it is measured. Scope divergence implies that there are different views about the set of relevant attributes that should be considered in an ESG rating. This is not avoidable, and perhaps even desirable given the various interpretations of the concept

measurement, and weights explain the difference between two ratings for a particular firm.

Third, we find that measurement divergence is in part driven by a *rater e ect*. This means that a firm that receives a high score in one category is more likely to receive high scores in all the other categories from that same rater. Similar effects have been shown in many other kinds of performance evaluations (see, e.g., Shrout and Fleiss (1979)). Our results hint at the existence of structural reasons for measurement divergence, including, for example, that ESG rating agencies usually divide labor among analysts by firm rather than by category.

Our methodology relies on two critical assumptions and we evaluate the robustness of each of them. First, indicators are assigned to categories based on our judgment. To evaluate the sensitivity of the results to this assignment, we also sorted the indicators according to a taxonomy provided by the Sustainability Accounting Standards Board (SASB).[2] The results based on this alternative taxonomy are virtually identical to those based on our assignment. Second, our linear aggregation rule is not industry-specific, while most ESG rating agencies use industry-specific aggregation rules. This approximation, however, seems to be relatively innocuous, since even a simple linear rule achieves a very high quality of fit. In addition to our analysis for 2014, the year that maximizes our sample size and includes KLD, we run a robustness check for the year 2017 without KLD and obtain very similar results.

We extend existing research that has documented the divergence of ESG ratings (Chatterji et al., 2016; Gibson et al., 2019). Our contribution is to explain why ESG ratings diverge by contrasting the underlying methodologies in a coherent framework and quantifying the sources of divergence. Our findings complement research documenting growing expectations from investors that companies take ESG issues seriously (Liang and Renneboog, 2017; Riedl and Smeets, 2017; Amel-Zadeh and Serafeim, 2018; Dyck et al., 2019). ESG ratings play an important role in translating these expectations into capital allocation decisions. Our research also provides an important empirical basis for future research on asset pricing and ESG ratings. In a theoretical model, Pastor et al. (2020) predict that investor preferences

thus essential to evaluate how changing investor expectations influence financial markets and corporate investments. Our study is also related to research on credit rating agencies (Bolton et al., 2012; Alp, 2013; Bongaerts et al., 2012; Jewell and Livingston, 1998), in the sense that we also investigate why ratings from different providers differ. Similar to Griffin and Tang (2011) and Griffin et al. (2013), we estimate the underlying rating methodologies to uncover how rating differences emerge.

The paper is organized as follows: Section 1 describes the data; Section 2 documents the divergence in the sustainability ratings from different rating agencies. Section 3 explains the way in which we develop the common taxonomy and estimate the aggregation procedures, while in Section 4 we decompose the overall divergence into the contributions of *scope*, *measurement*, and *weights* and document the rater effect. Finally, we conclude in Section 5 and highlight the implications of our findings.

1 Data

ESG ratings first emerged in the 1980s as a way for investors to screen companies not purely on financial characteristics but also on characteristics related 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, including faith-based organizations, the market for ESG ratings has widened dramatically, especially in the past decade. There are over 1,500 signatories to the Principles for Responsible Investing (PRI, 2018), who together own or manage over $80 trillion. As PRI signatories, these financial institutions commit to integrating ESG information into their investment decision-making. While growth in sustainable investing was initially driven by institutional investors, retail investors too are displaying an increasing interest, leading to substantial inflows for mutual funds that invest according to ESG criteria. Since ESG ratings

S&P Global bought RobecoSAM in 2019.

ESG rating agencies offer investors a way to screen companies for ESG performance in a similar way to how credit ratings allow investors to screen companies for creditworthiness. Yet despite this similarity there are at least three important differences between ESG ratings and credit ratings. First, while creditworthiness is relatively clearly defined as the probability of default, the definition of ESG performance is less clear. It is a concept based on values that are diverse and 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. There are competing reporting standards for ESG disclosure and almost none of the reporting is mandatory, giving corporations broad discretion regarding whether and what to report. Thus, ESG ratings provide a service to investors by collecting and aggregating information from across a spectrum of sources and reporting standards. These two differences serve to explain why the divergence between ESG ratings is so much more pronounced than the divergence between credit ratings, the latter being correlated at 99%.[3] And there is a third difference: ESG raters are paid by the investors who use them, not by the companies who get rated, as is the case with credit raters. As a result, the problem of ratings shopping, which has been discussed as a potential reason for credit ratings diverging (see, e.g., Bongaerts et al. (2012)) does not apply to ESG rating providers.

We use data from six different ESG rating providers: KLD[4], Sustainalytics, Vigeo Eiris, Asset4 (Refinitiv), MSCI, and RobecoSAM. Together, these providers represent most of the major players in the ESG rating space as reviewed in Eccles and Stroehle (2018) and cover a substantial part of the overall market for ESG ratings. 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 protocols of the indicators. We requested that the data set be as granular as possible.

Table 1 provides descriptive statistics of the aggregate ratings[5] and their sample char-

acteristics. The baseline year for our analysis is 2014, which is the year with the largest common sample when KLD is also included. Since most of the academic literature to date relies on KLD data, we think it is important to have it in our study. We also test whether our results are specific to the year of study, by rerunning the analysis for the year 2017 without KLD. As we show in the Internet appendix, the results are similar. Panel A of Table 1 shows the full sample, where the number of firms ranges from 1,665 to 9,662. Panel B of the same table shows the common sample of 924 firms. The mean and median ESG ratings are higher in the balanced sample for all providers, indicating that the balanced sample tends to drop lower-performing companies. Panel C shows the normalized common sample, in which ESG ratings are normalized in the cross-section to have zero mean and unit variance. Throughout the paper, we refer to these three samples as the full sample, the common sample, and the normalized sample.

2 Measurement of Divergence

To motivate our analysis, we illustrate the extent of divergence between the different rating agencies. First, we compute correlations between the ratings themselves as well as between inter-agency environmental, social, and governance dimensions. Second, we evaluate the heterogeneity of divergence across firms. Simple correlations, although easy to understand, can mask important heterogeneity in the data. To explore this, we analyze the mean absolute distance (MAD) to the average rating for each firm. Third, we explore disagreement in rankings. We illustrate that there is a very small set of firms that are consistently in the top or bottom quintile in all ratings. We then expand this approach to a thorough analysis for different quantiles using a simple statistic that we call the quantile ranking count (QRC).

2.1 Correlations of Aggregate Ratings

have the highest level of agreement between each other, with a correlation of 0.71. The correlations of the environmental dimension are slightly lower than the overall correlations, with an average of 0.53. The social dimension is on average correlated at 0.42, and the governance dimension has the lowest correlation, with an average of 0.30. KLD and MSCI clearly exhibit the lowest correlations with other raters, both for the rating and for the individual dimensions. These results are largely consistent with prior findings by Chatterji et al. (2016).

2.2 Heterogeneity in the Disagreement

Correlations may obscure firm level differences. For example, a weak correlation between two ratings can be driven either by similar disagreement for every firm or by extremely large disagreement for only a few firms. To analyze the heterogeneity of disagreement, we use the normalized common sample and compute the MAD to the average rating for each firm. Since the ratings have been normalized to have zero mean and unit variance, all values can be interpreted in terms of standard deviations. This yields a firm-specific measure of disagreement. Table 3 shows how the MAD measure is distributed and how it differs across sectors and regions. Panel A shows that the average MAD is 0.49, the median is 0.45, and the maximum 1.26, implying a slight positive skewness. Panels B and C of Table 3 show that there is no substantial variation across sectors or regions.

To illustrate what the rating disagreement looks like at the extremes, we focus on the 25 firms with the lowest and highest disagreement. Figure 1 shows the 25 firms with the lowest disagreement between raters. The average MAD for these 25 firms is 0.18. Among them, agreement is not perfect, but generally all rating agencies share a common view. Companies such as Amcor Limited, the Bank of Nova Scotia, and Heineken NV have high average ratings, and all six rating agencies tend to agree. For firms such as Morgan Stanley and Apple Inc., all raters agree tightly on scores in the middle range. Firms such as Amphenol Corporation, Intuitive Surgical Inc., and China Resources Land Ltd. have low average ratings, and all

few extreme observations. On average, Intel Corporation and GlaxoSmithKline have high ratings, Barrick Gold Corporation and AT&T Inc. have middle range ratings, and Porsche Automobil Holding and Philip Morris are among the worst rated. Yet in all cases, there is substantial disagreement around this assessment.

In summary, there is large heterogeneity in the level of disagreement across firms. Rating agencies agree on some firms, and disagree on others. There is, however, no obvious driver of this heterogeneity; it occurs for firms of all sectors and in all regions.

2.3 Quantile Analysis

Rankings 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 from the bottom quantile. With this approach, the disagreement in ratings would be less relevant than the disagreement in rankings.

Table 4 shows the firms that are in the top and bottom 20% of the common sample across all six raters. The first column in Table 4 provides an idea of how a sustainable investment portfolio that is based on a strict consensus of six rating agencies would have looked in 2014. There are only 15 companies that make it into the top 20% in all ratings, a small number considering that 20% of the sample equates to 184 companies. The second column of Table 4 lists the 23 companies that are included in the bottom 20% in all ratings. These are companies that one would expect to be consistently avoided by most sustainable investment funds.

The results presented in Table 4 are sensitive to the size of the chosen quantile. To provide a more general description of the divergence, we devise a measure that we call the *quantile ranking count*. First, we count how many firms are in the lower $q\%$ for every rating agency. We then calculate the ratio of this number to the total number of firms. If the ratings are perfectly aligned, then the exact same firms will be in the lower quantile ($q\%$). If the

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

In order to interpret the data, we simulate ratings with known and constant correlation. First, we simulate a random draw of 924×6 uniform realizations between the values of 0 and 1. We denote these realizations as $\epsilon_{k,f}$, where k is the rater and f is the index for the fictitious firm. Second, we create rankings for each rater and each firm as follows:

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

where the is calibrated to achieve an average correlation across all ratings. A value of = 0 implies that all the ratings are perfectly uncorrelated, and = 1 implies perfect correlation. We calibrate the to achieve an average correlation of 10, 20,..., 80, 90, and 95%. Finally, from the simulated data we computed the quantile ranking counts (QRCs) for each quantile q. We run this simulation a thousand times and take the average of each data point.

In Figure 3 we present the quantile ranking count for the overall ESG rating for all rating agencies and firms in the common sample.[6] The thick orange line indicates the counts of the actual data and the dashed gray lines reflect the counts of the simulated data. We begin by observing the 20% quantile, which corresponds to the case shown in Table 4. In Figure 3, the thick line is situated between the fourth and the fifth gray lines. This corresponds to an implied correlation of between 80 and 70%. 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 have correlations of between 70% and 80%. At the 50% quantile the thick line crosses the line that corresponds to the 70% implied correlation. Finally, at the 90% quantile the implied correlation is close to 40%. This indicates that there is more disagreement among the top rated firms.

In summary, this section has established the following stylized facts about ESG rating

for firms that are ranked near the top of the distribution. As a result, it is likely that portfolios that are based on different ESG ratings have substantially different constituents, and portfolios that are restricted to top performers in all ratings are extremely constrained to very few eligible companies.

3 Taxonomy and Aggregation Rules

ESG ratings are indices that aggregate a varying number of indicators into a score that is designed to measure a firm's ESG performance. Conceptually, such a rating can be described in terms of scope, measurement, and weights. Scope refers to the set of attributes that describe a company's ESG performance. Measurement refers to the indicators that are used to produce a numerical value for each attribute. Weights refers to the function that combines multiple indicators into one rating. Figure 4 provides an illustration of this schematic view.

The three elements—scope, measurement, and weights—translate into three distinct sources of divergence. Scope divergence results when two raters use a different set of attributes. For example, all rating agencies in our sample consider a firm's water consumption, but only some include a firm's lobbying activities. Measurement divergence results when two raters use different indicators to measure the same attribute. For instance, the attribute of gender equality could be measured by the percentage of women on the board, or by the gender pay gap within the workforce. Both indicators are a proxy for gender equality, but they are likely to result in different assessments. Finally, weights divergence[7] results when raters use different aggregation functions to translate multiple indicators into one ESG rating. The aggregation function could be a simple weighted average, but it could also be a more complex function involving non-linear terms or contingencies on additional variables such as industry affiliation. A rating agency that is more concerned with GHG Emissions than Electromagnetic Fields will assign different weights than a rating agency that cares equally about both issues. Differences in the aggregation function lead to different ratings,

3.1 Taxonomy

The goal of this paper is to decompose the overall divergence between ratings into the sources of scope, measurement, and weights. This is not trivial, because at the granular level the approach of each rating agency looks very different. Each rater chooses to break down the concept of ESG performance into different indicators, and organizes them in different hierarchies. For example, at the first level of disaggregation, Vigeo Eiris, RobecoSAM, MSCI, 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 38 and 282 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.

In order to perform a meaningful comparison of these different rating systems, we impose our own taxonomy on the data, as shown in Table 5. We develop this taxonomy using a bottom-up approach. First, we create 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 contains 709 indicators. Second, we group indicators that describe the same attribute in the same *category*. For example, we group together all indicators related to resource consumption or those related to community relationships. Third, we iteratively refine the taxonomy, following two rules: (a) each indicator is assigned to only one category, and (b) a new category is established when at least two indicators from different raters both describe an attribute that is not yet covered by existing categories. The decision is purely based on the attribute that indicators intend to measure, regardless of the method or data source that is used. For example, indicators related to Forests were taken out of the larger category of Biodiversity to form

KLD, RobecoSAM, and MSCI have 78, 80, and 68, respectively, and Vigeo Eiris has 38. Some categories—Forests, for example—contain just one indicator from two raters. Others, such as Supply Chain, contain several indicators from all raters. Arguably, Forests is much more narrow a category than Supply Chain. The reason for this difference in broadness is that there were no indicators in Supply Chain that together represented a more narrow 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.

Table 5 already reveals that there is considerable scope divergence. On the one hand, there are categories that are considered by all six raters, indicating some sort of lowest common denominator of categories that are included in an ESG rating. These are Biodiversity, Employee Development, Energy, Green Products, Health and Safety, Labor Practices, Product Safety, Remuneration, Supply Chain, and Water. On the other hand, there are many empty cells, which shows that far from all categories are covered by all ratings. There are gaps not only for categories that could be described as specialized, such as Electromagnetic fields, but also for the category Taxes, which could be viewed as 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 six raters. Asset4 has, with 42, the most unclassified indicators, almost all of which stem from Asset4's economic dimension. This dimension contains indicators such as net income growth or capital expenditure, which are not considered by any other rating agency. MSCI has 34 unclassified indicators; these come from so-called exposure scores, which MSCI has as a counterpart to most of their management scores. These exposure scores are a measure of how important or material the category is for the specific company. None of the other raters have indicators that explicitly measure such exposure.

The taxonomy imposes a structure on the data that allows a systematic comparison.

alternative taxonomy as a robustness check. 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 (SASB). SASB has identified 26 so-called general issue categories, which are the results of a comprehensive stakeholder consultation process. As such, these categories represent the consensus of a wide range of investors and regulators on the scope of relevant ESG categories. We map all indicators against these 26 general issue categories, again requiring that each indicator can only be assigned to one category. This alternative taxonomy, along with results that are based on it, is provided in the Internet appendix. All our results hold also for this alternative taxonomy.

3.2 Category Scores

On the basis of our taxonomy, we can study measurement divergence by comparing the assessments of different raters at the level of categories. To do so, we create category scores (C) for each category, firm, and rater. Category scores are 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 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	$(1, 6)$
F	Firms	f	$(1, 924)$

The category score is computed as

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. Some rating agencies employ different sets of indicators for different industries. Such industry-specific considerations about measurement are also reflected in the category scores, since they take the average of all indicator values that are available.

Table 6 shows the correlations between the categories. The correlations are calculated on the basis of complete pairwise observations per category and rater pair. They range from -0.5 for Responsible Marketing between KLD and Sustainalytics to 0.92 for Global Compact Membership between Sustainalytics and Asset4. When comparing the different rater pairs, KLD and MSCI have the highest average correlation, with 0.69, whereas all other ratings have relatively low correlations with KLD, ranging from 0.12 to 0.21.

Beyond these descriptive observations Table 6 offers two insights. First, correlation levels are heterogeneous. Environmental Policy, for instance, has an average correlation level of 0.55. This indicates that there is at least some level of agreement regarding the existence and quality of the firms' environmental policy. But even categories that measure straightforward facts that are easily obtained from public records do not all have high levels of correlation. Membership of the UN Global Compact and CEO/Chairperson separation, for instance, show correlations of 0.92 and 0.59, respectively. Health and Safety is correlated at 0.30, Taxes at 0.04. There are also a number of negative correlations, such as Lobbying between Sustainalytics and Vigeo Eiris or Indigenous Rights between Sustainalytics and Asset4. In these cases, the level of disagreement is so severe that rating agencies reach not just different, but opposite conclusions.

The second insight is that correlations tend to increase with granularity. For example, the correlations of the categories Water and Energy are on average 0.36 and 0.38, respectively.

disagreement on individual categories cancels out during aggregation. It may also be the case that rating agencies assess a firm relatively strictly in one category and relatively leniently in another. A concern might be that the low correlations at the category level result from misclassification in our taxonomy, in the sense that highly correlated indicators were sorted into different categories. While we cannot rule this out completely, the alternative taxonomy based on SASB criteria mitigates this concern. It is a much less granular classification, which therefore should decrease the influence of any misclassification. The average correlation per rater pair, however, hardly changes when using this alternative taxonomy. This provides reassurance that the observed correlation levels are not an artefact of misclassification in our taxonomy[8].

In sum, this section has shown that there is substantial measurement divergence, indicated by low levels of correlations between category scores. Furthermore, the section has revealed that measurement divergence is heterogeneous across categories. Next, we turn to weights divergence, the third and final source of divergence.

3.3 Aggregation Rule Estimation

Based on the category scores we can proceed with an analysis of weights divergence. To do so, we estimate the aggregation rule that transforms the category scores C_{fkj} into the rating R_{fk} for each rater k. It turns out that a simple linear function is sufficient. We perform a non-negative least squares regression and present the resulting category weights in Table 7. In addition, we perform several robustness checks that relax assumptions related to linearity, and we explore the sensitivity to using alternative taxonomies and data from a different year.

Category scores, as defined in Section 3.2, serve as independent variables. When there are no indicator values available to compute the category score for a given firm the score is set to zero. This is necessary in order to run regressions without dropping all categories with

indicator is treated as a separate rater-specific category.

We perform a non-negative least squares regression, which includes the constraint that coefficients cannot be negative. This reflects the fact that we know a priori the directionality of all indicators, and can thus rule out negative weights in a linear function. Thus, we estimate the weights (w_{kj}) with the following specification:

$$R_{fk} = \sum_{j \in 1,m)} C_{fkj} \times w_{kj} + \epsilon_{fk}$$
$$w_{kj} \geq 0.$$

Since all the data has been normalized, we exclude the constant term. Due to the non-negativity constraint we calculate the standard errors by bootstrap. We focus on the R^2 as a measure of quality of fit.

The results are shown in Table 7. MSCI has the lowest R^2, with 0.79. Sustainalytics the second lowest, with 0.90. The regressions for KLD, Vigeo Eiris, Asset4, and RobecoSAM have R^2 values of 0.99, 0.96, 0.92, and 0.98, respectively. The high R^2 values indicate that a linear model based on our taxonomy is able to replicate the original ratings quite accurately.

The regression represents a linear approximation of each rater's aggregation rule, and the regression coefficients can be interpreted as category weights. Since all variables have been normalized, the magnitude of the coefficients is comparable and indicates the relative importance of a category. Most coefficients are highly significant. There are some coefficients that are not significant at the 5% threshold, which means that our estimated weight is uncertain. Those coefficients are, however, much smaller in magnitude in comparison to the significant coefficients; in fact most of them are close to zero and thus do not seem to an important influence the aggregate ESG rating.

There are substantial differences in the weights for different raters. For example, the three most important categories for KLD are Climate Risk Management, Product Safety

categories that have zero weight for all raters, such as Clinical Trials and Environmental Fines, GMOs, and Ozone-Depleting Gases. This suggests these categories have no statistical relevance for any of the aggregate ratings. These observations highlight that different raters have substantially different views about which categories are most important. In other words, there is substantial weights divergence between raters.

The estimation of the aggregation function entails several assumptions. To ensure the robustness of our results, we evaluate several other specifications. The results of these alternative specifications are summarized in Table 8. None of them offer substantial improvements in the quality of fit over the non-negative linear regression.

First, we run an ordinary least squares regression in order to relax the non-negativity constraint. Doing so leads only to small changes and does not improve the quality of fit for any rater. Second, we run neural networks in order to allow for a non-linear and flexible form of the aggregation function. As neural networks are prone to overfitting, we report the out-of-sample fit. We randomly assign 10% of the firms to a testing set, and the rest to a training set.[9] To offer a proper comparison, we compare their performance to the equivalent out-of-sample R^2 for the non-negative least squares procedure. We run a one-hidden-layer neural network with a linear activation function and one with a relu activation function. Both perform markedly better for MSCI, but not for any of the other raters. This implies that the aggregation rule of the MSCI rating is to some extent non-linear. In fact, the relatively simple explanation seem to be industry-specific weights. In unreported tests, we confirm that the quality of fit for MSCI is well above 0.90 in industry sub-samples even for a linear regression. Third, we implement a random forest estimator as an alternative non-linear technique. This approach yields, however, substantially lower R^2 values for most raters.

We also check whether the taxonomy that we imposed on the original indicators has an influence on the quality of fit. To this end, we replicate the non-negative least squares estimation of the aggregation rule using the SASB taxonomy.[10] The quality of fit is virtually

the most notable change being a small increase of 0.03 for the MSCI rating. Finally, we perform the regression using data from the year 2017 (without KLD) instead of 2014. In this case, the quality of fit is worse for MSCI and Asset4, indicating that their methodologies have changed over time. In sum, we conclude that none of the alternative specifications yields substantial improvements in the quality of fit over the non-negative least squares model.

4 Decomposition and Rater E ect

So far, we have shown that scope, measurement, and weights divergence exist. In this section, we aim to understand how these sources of divergence together explain the divergence of ESG ratings. Specifically, we decompose ratings divergence into the contributions of scope, measurement, and weights divergence. We also investigate the patterns behind measurement divergence and detect a rater effect, meaning that measurement is influenced by the rating agency's general view of the rated firm.

4.1 Scope, Measurement, and Weights divergence

We develop two alternative approaches for the decomposition. First, we arithmetically decompose the difference between two ratings into differences due to scope, due to measurement, and due to weights. This approach identifies exactly the shift caused by each source of divergence. Yet as these shifts are not independent of each other, the approach is not ideal to determine their relative contribution to the total divergence. Thus, in a second approach, we adopt a regression-based approach to provide at least a range of the relative contributions of scope, measurement, and weights.

4.1.1 Arithmetic Decomposition

The arithmetic variance decomposition relies on the taxonomy, the category scores, and

in one of the two ratings. Measurement divergence is isolated by calculating both ratings with identical weights, so that differences can only stem from differences in measurement. Weights divergence is what remains of the total difference.

Let $\hat{R}_{fk}$ (where $k \in a, b$) be the rating provided by rating agency a and rating agency b for a common set of f companies. $\hat{R}_{fk}$ denotes the fitted rating and $\hat{w}_{kj}$ the estimated weights for rater k and category j based on the regression in Table 7. Thus, the decomposition is based on the following relationship:

$$\hat{R}_{fk} = C_{fkj} \times \hat{w}_{kj} \tag{4}$$

Common categories, which are included in the scope of both raters, are denoted as $C_{fkj_{com}}$. Exclusive categories, which are included by only one rater, are 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). Similarly, $\hat{w}_{aj_{com}}$ and $\hat{w}_{bj_{com}}$ are the weights used by rating agencies a and b for the common categories, and $\hat{w}_{aj_{a,ex}}$ are the weights for the categories only measured by a, and analogously $\hat{w}_{bj_{b,ex}}$ for b. We separate the rating based on common and exclusive categories as follows:

Definition 2 ***Common and Exclusive Categories***

For $k \in \{a, b\}$ define:

$$\begin{aligned} \hat{R}_{fk,com} &= C_{fkj_{com}} \times \hat{w}_{kj_{com}} \\ \hat{R}_{fk,ex} &= C_{fkj_{k,ex}} \times \hat{w}_{kj_{k,ex}} \\ \hat{R}_{fk} &= \hat{R}_{fk,com} + \hat{R}_{fk,ex} \end{aligned} \tag{5}$$

On this basis, we can provide terms for the contributions of scope, measurement, and weights divergence to the overall divergence.

Definition 3 ***Scope Measurement and Weights***

$$
\begin{aligned}
_{scope} &= C_{faj_{a\ ex}} \times \hat{w}_{aj_{a\ ex}} \quad C_{fbj_{b\ ex}} \times \hat{w}_{bj_{b\ ex}} \\
_{meas} &= (C_{faj_{com}} \quad C_{fbj_{com}}) \times \hat{w} \\
_{weights} &= C_{faj_{com}} \times (\hat{w}_{aj_{com}} \quad \hat{w}\ ) \quad C_{fbj_{com}} \times (\hat{w}_{bj_{com}} \quad \hat{w}\ )
\end{aligned} \tag{7}
$$

where $\hat{w}^*$ are the estimates from pooling regressions using the common categories

$$
\begin{pmatrix} \hat{R}_{fa,com} \\ \hat{R}_{fb,com} \end{pmatrix} = \begin{pmatrix} C_{faj_{com}} \\ C_{fbj_{com}} \end{pmatrix} \times w^* + \begin{pmatrix} \epsilon_{fa} \\ \epsilon_{fb} \end{pmatrix} \tag{8}
$$

Scope divergence $_{scope}$ is the difference between ratings that are calculated using only mutually exclusive categories. Measurement divergence $_{meas}$ is calculated based on the common categories and identical weights for both raters. Identical weights $\hat{w}^*$ are estimated in equation 8, which is a non-negative pooling regression of the stacked ratings on the stacked category scores of the two raters. Since the least squares make sure that we maximize the fit with $\hat{w}^*$, we can deduce that $_{meas}$ captures the differences that are exclusively due to differences in the category scores. Weights divergence ($_{weights}$) is simply the remainder of the total difference, or, more explicitly, a rater's category scores multiplied with the difference between the rater-specific weights $\hat{w}_{aj_{com}}$ and $\hat{w}^*$. It must be noted that all these calculations are performed using the fitted ratings $\hat{R}$ and the fitted weights $\hat{w}$, since the original aggregation function is not known with certainty.

By way of an example, Figure 5 shows the decomposition of the rating divergence between ratings from Asset4 and KLD for Barrick Gold Corporation. The company received a normalized rating of 0.52 from Asset4 vs. -1.10 from KLD. The resulting difference of 1.61 is substantial considering that the rating is normalized to unit variance. The difference between our fitted ratings is slightly lower, at 1.60, due to residuals of +0.09 and +0.10 for the fitted ratings of Asset4 and KLD, respectively. This difference consists of 0.41 scope divergence, 0.77 measurement divergence, and 0.42 weights divergence. The three most relevant categories that contribute to scope divergence are Taxes, Resource Efficiency, and Board,

muneration than Asset4, but a higher score for Indigenous Rights. The different assessment of Remuneration accounts for about a third of the overall rating divergence. The most relevant categories for weights divergence are Community and Society, Biodiversity, and Toxic Spills. Different weights for the categories Biodiversity and Toxic Spills drive the two ratings apart, while the weights of Community and Society compensate part of this effect. The combined effect of the remaining categories is shown for each source of divergence under the label Other". This example offers a concrete explanation of why these two specific ratings differ.

Cross-sectional results of the decomposition are presented in Table 9, where Panel A shows the data for each rater pair and Panel B shows averages per rater based on Panel A. The first three columns show the mean absolute values of scope, measurement, and weights divergence. The column Fitted" presents the difference between the fitted ratings $\hat{R}_{fk_1} \quad \hat{R}_{fk_2}$, and the column True" presents the difference between the original ratings $|R_{fk_1} \quad R_{fk_2}|$. Since the ratings have been normalized to have zero mean and unit variance, all values can be interpreted in terms of standard deviations.

Panel A shows that, on average, measurement divergence is the most relevant driver of ESG rating divergence, followed by scope divergence and weights divergence. Measurement divergence causes an average absolute shift of 0.54 standard deviations, ranging from 0.39 to 0.67. Scope divergence causes an average absolute shift of 0.48 standard deviations, ranging from 0.19 to 0.86. Weights divergence causes an average absolute shift of 0.34 standard deviations, ranging from 0.11 to 0.57. The fitted rating divergence is similar to the true rating divergence, which corresponds to the quality of fit of the estimations in Section 3.3.

Panel B highlights differences between raters. MSCI is the only rater where scope instead of measurement divergence causes the largest shift. With a magnitude of 0.85, the scope divergence of MSCI is twice as large as the scope divergence of any other rater. MSCI is also the only rater for which weights divergence is almost equally as relevant as measurement divergence. KLD is noteworthy in that it has the highest value for measurement divergence

scope divergence of MSCI with respect to all other raters.

The sum of scope, measurement, and weights divergence exceeds the overall rating divergence in all cases. This suggests that the three sources of divergence are negatively correlated and partially compensate each other. This is not surprising given that, by their construction, measurement and weights divergence are related through the estimation of $\hat{w}$. While the arithmetic decomposition is exact for any given firm, the averages of the absolute distances do not add up to the total. Nevertheless, the decomposition shows that on average measurement divergence tends to cause the greatest shift, followed by scope divergence, and finally weights divergence.

4.1.2 Regression-Based Decomposition

In this section we present an alternative decomposition methodology. We regress the ratings of one agency on the ratings of another, and analyze the gain in explanatory power that is due to variables representing scope, measurement, and weights divergence. Doing so addresses the key shortcoming of the methodology from the previous section, that the three sources of divergence do not add up to the total divergence.

Definition 4 ***Measurement Scope and Weights Variables***

$$Scope_{fa,b} = C_{fbj_{b,ex}} \cdot \hat{w}_{bj_{b,ex}} \tag{9}$$

$$Meas_{fa,b} = C_{fbj_{com}} \cdot \hat{w}_{aj_{com}} \tag{10}$$

$$Weight_{fa,b} = C_{faj_{com}} \cdot \hat{w}_{bj_{com}} \tag{11}$$

Similar to the prior decomposition, this approach also relies on the taxonomy, category scores, and the weights estimated in Section 3.3. $Scope_{fa,b}$ consists of only the categories and the corresponding weights that are exclusive to rater b. $Meas_{fa,b}$ consists of the category scores in rater b and rater a's corresponding weights for the common categories. Finally, the

$$\hat{R}_{fb} = \beta \cdot \hat{R}_{fa} + \beta_s \cdot Scope_{fa,b} + \beta_m \cdot Meas_{fa,b} + \beta_w \cdot Weight_{fa,b} + \epsilon \tag{12}$$

The fitted rating $\hat{R}_{fb}$ is the outcome of the dot product between the category scores C_{fbj} and rater b's estimated weights $\hat{w}_{bj}$; the equivalent is true 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 weights.

Given that the three terms scope, measurement, and weights are correlated with each other, the order in which we add them as regressors to regression 12 matters. We thus run partialing-out regressions in order to calculate a lower and an upper bound of the additional explanatory power 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 between the R^2 values. By changing the regressors in the baseline, the contribution of scope changes. We therefore run regressions in all possible combinations. For example, for scope we estimate the following eight regressions:

$$\begin{aligned}
\hat{R}_{fb} &= \cdot \hat{R}_{fa} && && && + \epsilon \implies R^2 \\
\hat{R}_{fb} &= \cdot \hat{R}_{fa} + {}_s \cdot Scope_{fa\,b} && && && + \epsilon_1 \implies R_1^2 \\
\hat{R}_{fb} &= \cdot \hat{R}_{fa} && + {}_m \cdot Meas_{fa\,b} && && + \epsilon_2 \implies R_2^2 \\
\hat{R}_{fb} &= \cdot \hat{R}_{fa} + {}_s \cdot Scope_{fa\,b} && + {}_m \cdot Meas_{fa\,b} && && + \epsilon_3 \implies R_3^2 \\
\hat{R}_{fb} &= \cdot \hat{R}_{fa} && && + {}_w \cdot Weight_{fa\,b} && + \epsilon_4 \implies R_4^2 \\
\hat{R}_{fb} &= \cdot \hat{R}_{fa} + {}_s \cdot Scope_{fa\,b} && && + {}_w \cdot Weight_{fa\,b} && + \epsilon_5 \implies R_5^2 \\
\hat{R}_{fb} &= \cdot \hat{R}_{fa} && + {}_m \cdot Meas_{fa\,b} && + {}_w \cdot Weight_{fa\,b} && + \epsilon_6 \implies R_6^2
\end{aligned}$$

The results of the statistical decomposition are presented in Table 10, where Panel A shows the data for each rater pair and Panel B shows averages per rater. The first column presents the baseline R^2, which, in the first row, for example, is simply regressing the KLD rating on the Vigeo Eiris rating. The second column is the R^2 from a regression that includes all four covariates—that is, it includes rating a plus the scope, measurement, and weights variables. The next six columns indicate the minimum and maximum R^2 gain of explanatory power due the inclusion of the scope, measurement, and weights variables.

The first column shows that the average explanatory power when trying to simply explain one rating with another is 0.34 and fluctuates between 0.16 and 0.56. The second column shows that when including the terms for scope, measurement, and weights, the R^2 rises on average to 0.84 (ranging from 0.44 to 0.97). Thus, the additional variables improve the fit by 0.51 on average. Scope offers the greatest improvement in explanatory power, with an average minimum gain of 0.14 and an average maximum gain of 0.35. This is almost equal to measurement with an average gain of at least 0.14 and at most 0.35. The addition of weights leads to far lower gains, of at least 0.01 and at most 0.04. These ranges indicate the relative contribution of the three sources of divergence to the total divergence.

The findings are consistent with the results from the previous decomposition. While scope divergence is slightly more relevant than measurement divergence in this decomposition, the two are clearly the dominant sources of divergence. Weights divergence is less relevant in explaining the rating divergence. Looking at specific raters in Panel B also reaffirms the prior finding that scope divergence is much more relevant for MSCI than for any other rater. Asset4 has the lowest values for scope divergence, which is also consistent with the previous results. In sum, scope and measurement divergence are the predominant sources of ESG rating divergence, with weights divergence playing a minor role in comparison.

4.2 Rater E ect

One could argue that measurement divergence is the most problematic source of diver-

The rater effect describes a sort of bias, where performance in one category influences perceived performance in other categories. This phenomenon has been extensively studied in sociology, management, and psychology, especially in performance evaluation (see Shrout and Fleiss (1979)). The process of evaluating firms' ESG attributes seems prone to a rater effect. Evaluating firm performance in the categories Human Rights, Community and Society, Labor Practices, etc. requires rating agencies to use some degree of judgment. The rater effect implies that when the judgement of a company is positive for one particular indicator, it is also likely to be positive for another indicator. We evaluate the rater effect using two procedures. First, we estimate fixed effects regressions comparing categories, firms, and raters. Second, we run rater-specific LASSO regressions to evaluate the marginal contribution of each category.

4.2.1 Rater Fixed E ects

The first procedure is based on simple fixed effects regressions. A firm's category scores depend on the firm itself, on the rating agency, and on the category being rated. We examine to what extent those fixed effects increase explanatory power in the following set of regressions:

$$C_{fkj} = {}_f\mathbb{1}_f + \epsilon_{fkj,1} \tag{13}$$

$$C_{fkj} = {}_f\mathbb{1}_f + \gamma_{fk}\mathbb{1}_{f\ k} + \epsilon_{fkj,2} \tag{14}$$

$$C_{fkj} = {}_f\mathbb{1}_f + \gamma_{fj}\mathbb{1}_{f\ j} + \epsilon_{fkj,3} \tag{15}$$

$$C_{fkj} = {}_f\mathbb{1}_f + \gamma_{fk}\mathbb{1}_{f\ k} + \gamma_{fj}\mathbb{1}_{f\ j} + \epsilon_{fkj,4} \tag{16}$$

where $\mathbb{1}_f$ are dummies for each firm, $\mathbb{1}_{f\ k}$ is an interaction term between firm and rater fixed effects, and $\mathbb{1}_{f\ j}$ is an interaction term between firm and category fixed effects. The vector C_{fkj} stacks all cross-sectional scores for all common categories across all raters. We drop pure category and rater fixed effects because of the normalization at the rating and

The increment in R^2 between the two regression is the rater effect. The third and fourth regressions repeat the procedure, but with the additional inclusion of category-firm fixed effects. The results of these regressions are shown in Table 11.

We detect a clear rater effect. Firm dummies alone explain 0.22 of the variance of the scores in equation 13. When including firm-rater dummies, however, the R^2 increases to 0.38, an addition of 0.16. Similarly, the difference in R^2 between equation 15 and equation 16 yields an increase of 0.15. The rater effect therefore explains about 0.15 to 0.16 of the variation in category scores. The rater effect is relevant in comparison to the other dummies. Comparing the estimates of equations 15 and 13, we find that including firm-category dummies improves the fit by 0.25. Similarly, comparing the outcomes of regressions 16 and 14 yields an increase of 0.24. Thus, firm dummies explain 0.22, firm-category dummies 0.24-0.25, and firm-rater dummies 0.15-0.16. Even though the rater effect is smaller than the other two, it has a substantial influence on the category scores.

4.2.2 A LASSO Approach to the Rater E ect

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 categories are 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 estimating the linear aggregation rules with a LASSO regression. The LASSO estimator 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 combination of regressors that

increases, the variables with the smallest explanatory power are eliminated. In other words, the first category that has the smallest marginal contribution to the 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.

Table 12 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. 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 later categories only contribute to the R^2 by their orthogonal component.

In the computation of the aggregation rules (Table 7), the number of categories including the unclassified indicators covered by Vigeo Eiris, RobeccoSAM, Asset4, KLD, MSCI, and Sustainalytics are 28, 45, 95, 41, 61, and 63, respectively. Therefore, 10% of the possible regressors are 3, 5, 10, 4, 6, and 6, respectively. We have highlighted these fields in Table 12. Hence, 10% of the categories explain more than a fifth (0.21) of the variation in Vigeo Eiris's ratings, and this figure is 0.75 for RobeccoSAM, 0.63 for Asset4, 0.23 for KLD, 0.46 for Sustainalytics, and only 0.13 for MSCI. This illustrates the presence of a rater effect.

For completeness, in Figure 6 we present the increase in the R^2 for each rating agency for all categories. The curves reflect the evolution of the R^2. The last part of the curve to the right coincides with an unrestricted OLS estimate where all variables are included. These figures provide the same message we obtained from observing the R^2 before. KLD and MSCI have the smallest cross-category correlation, judging by the slope in Figure 6(a) and 6(f). Sustainalytics is the second flattest, followed by Vigeo Eiris and Asset 4, thus leaving RobecoSAM as the rating agency where just a few categories already explain most of the ESG rating.

The rater effect of ESG rating agencies establishes an interesting parallel to finance

Extending from the fact that there are biases in credit ratings, a lot of emphasis in the literature has been on understanding the structural drivers of such biases (see, e.g., Bolton et al. (2012); Bongaerts et al. (2012); Alp (2013)). This suggests a future avenue of research could be to also understand what drives the rater effect of ESG rating agencies, and whether incentive structures play a role.

A potential explanation for the rater effect is that rating agencies are mostly organized in such a way that analysts specialize in firms rather than indicators. A firm that is perceived as good in general may be seen through a positive lens and receive better indicator scores than a firm that is perceived as bad in general. In discussions with 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, a firm's willingness to disclose might also explain parts of the rater effect.

5 Conclusions

The contribution of this article is to explain why ESG ratings diverge. We develop a framework that allows a structured comparison of very different rating methodologies. This allows us to separate the difference between ratings into the components scope, measurement, and weights divergence. We find that measurement divergence is the most important reason why ESG ratings diverge, i.e. different raters measure the performance of the same firm in the same category differently. Human Rights and Product Safety are categories for which such measurement disagreement is particularly pronounced. Slightly less important is scope divergence, i.e. raters consider certain categories that others do not. For example, a company's lobbying activities are considered only by two out of the six raters in our sample. The least important type of divergence is weights divergence, i.e. disagreement about the

is not only due to random measurement error, but is partly driven by some form of rater-specific bias. This also implies that some ESG ratings could be replicated with a reduced set of categories, since category assessments are partially redundant in a statistical sense. Although we do not conclusively identify the cause of the rater effect, one possible explanation is that ESG rating agencies divide analyst labor by firm and not by category, so that an analyst's overall view of a company could propagate into the assessments in different categories. Further research is, however, needed to fully understand the reasons behind the rater effect.

Our findings demonstrate that ESG rating divergence is not merely driven by differences in opinions, but also by disagreements about underlying data. Scope and weights divergence both represent disagreement about what the relevant categories of ESG performance are, and how important they are relative to each other. It is legitimate that different raters take different views on these questions. In fact, a variety of opinions may be desirable given that the users of ESG ratings also have heterogeneous preferences for scope and weights. In particular, different investors will hold different views regarding which categories they deem material—that is to say, relevant for a firm's business success. Measurement divergence is problematic, however, if one accepts the view that ESG ratings should ultimately be based on objective observations that can be ascertained.

Our results have important implications for researchers, investors, companies, and rating agencies. Researchers should carefully choose the data that underlies future studies involving ESG performance. Certain results that have been obtained on the basis of one ESG rating might not be replicable with the ESG ratings of another rating agency. In particular, our results indicate that divergence is very pronounced for KLD—the data on which the majority of existing academic research into ESG has been based. Basically, researchers have three options when it comes to dealing with the divergence of ESG ratings. One is to include several ESG ratings in the analysis (see, e.g., Liang and Renneboog (2017)). This is reasonable when the intention is to measure consensus ESG performance" as it is perceived by financial

their study. Third, researchers can construct hypotheses around attributes that are more narrowly defined than ESG performance, and rely on verifiable and transparent measures of, for instance, GHG Emissions or Labor Practices. In such a case, it would still be important to consider several alternative measures in order to avoid uncertainty in measurement. But at least uncertainty around the selection and the weighting of different categories would be avoided.

Turning to investors, our methodology enables them to understand why a company has received different ratings from different rating agencies. The example in Figure 5 illustrates how a company can disentangle the various sources of divergence and trace down to specific categories. For instance, investors could reduce the discrepancy between ratings by obtaining indicator-level data from several raters and imposing their own scope and weights on the data. The remaining measurement divergence could be traced to the indicators that are driving the discrepancy, potentially guiding an investor's additional research. Averaging indicators from different providers is an easy way to eliminate measurement divergence as well—the rater effect suggests, however, that this approach may be problematic because the discrepancies are not randomly distributed. Alternatively, investors might rely on one rating agency, after convincing themselves that scope, measurement, and weights are aligned with their objectives.

For companies, our results highlight that there is substantial disagreement about their ESG performance. This divergence occurs not only at the aggregate level but is actually even more pronounced in specific sub-categories of ESG performance, such as Human Rights or Energy. This situation presents a challenge for companies, because improving scores with one rating provider will not necessarily result in improved scores at another. Thus, ESG ratings do not, currently, play as important a role as they could in guiding companies toward improvement. To change this situation, companies should work with rating agencies to establish open and transparent disclosure standards, and ensure that the data that they themselves disclose is publicly accessible.

academics, to evaluate and cross-check the agencies' measurements. Finally, rating agencies should seek to understand what drives the rater effect, in order to avoid potential biases. By taking these steps, rated firms would have clearer signals about what is expected of them, and investors could determine more precisely whether ESG ratings are aligned with their objectives.

6 References

A. Alp. Structural Shifts in Credit Rating Standards. *Journal of Finance*, 68(6):2435–2470, 2013.

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.

P. Bolton, X. Freixas, and J. Shapiro. The Credit Ratings Game. *Journal of Finance*, 67 (1):85–111, 2012.

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.

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.

A. Dyck, K. V. Lins, L. Roth, and H. F. Wagner. Do institutional investors drive corporate social responsibility? International evidence. *Journal of Financial Economics*, 131(3): 693–714, 2019.

R. G. Eccles and J. C. Stroehle. Exploring Social Origins in the Construction of ESG Measures. 2018.

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

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.

S. M. Hartzmark and A. B. Sussman. Do Investors Value Sustainability? A Natural Experiment Examining Ranking and Fund Flows. *The Journal of Finance*, 74:2789–2837, 2019.

H. Hong and M. Kacperczyk. The price of sin: The effects of social norms on markets. *Journal of Financial Economics*, 1(93):15–36, 2009.

H. Hong and L. Kostovetsky. Red and blue investing: Values and finance. *Journal of Financial Economics*, 103(1):1–19, 2012.

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. *The Review of Financial Studies*, 33(3):1067–1111, 2020.

H. Liang and L. Renneboog. On the foundations of corporate social responsibility. *Journal of Finance*, 72(2):1–59, 2017.

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.

L. Pastor, R. F. Stambaugh, and L. A. Taylor. Sustainable Investing in Equilibrium. SSRN

H. Servaes. The Impact of Corporate Social Responsibility on Firm Value : The Role of Customer Awareness. *Management Science*, 1909:1–17, 2013.

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

Figure 1
Firms with low disagreement

Normalized ratings for the 25 firms with the lowest mean absolute distance to the average rating (MAD) within the normalized common sample (n=924). Firms are sorted by their average rating. Each rating agency is plotted in a different color.

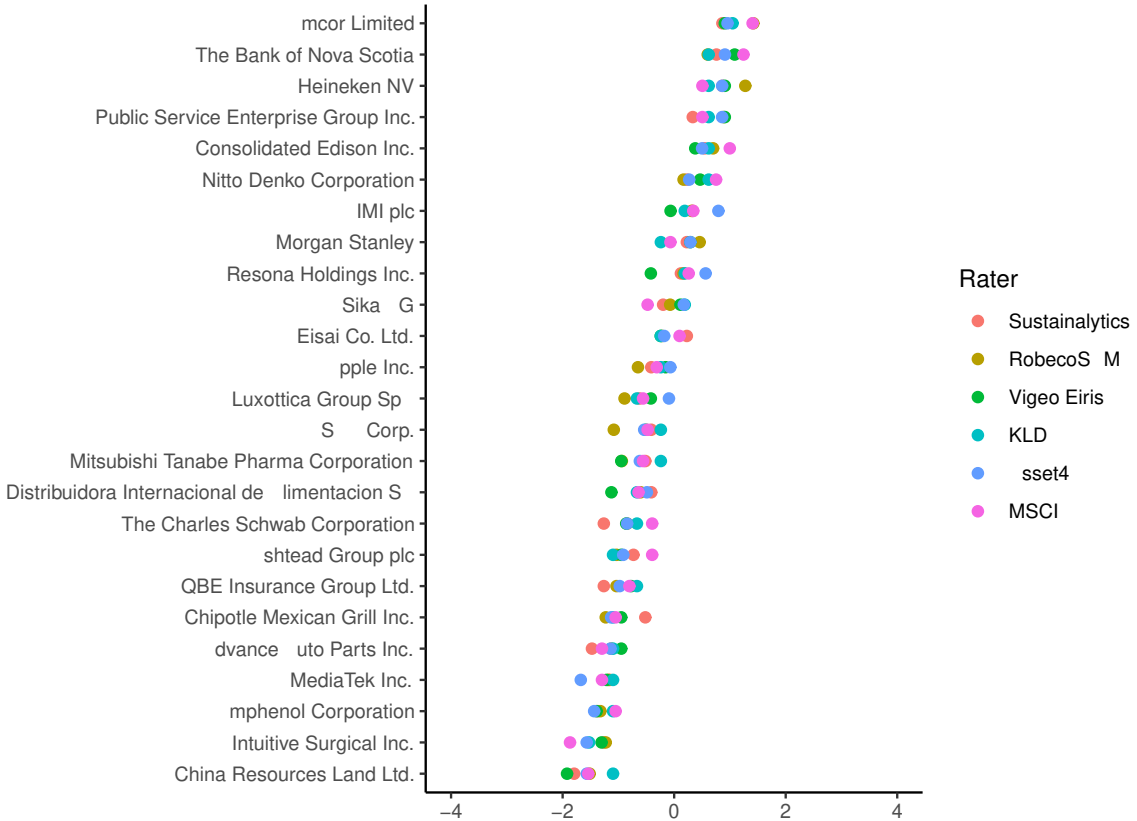


Figure 2
Firms with high disagreement

Normalized ratings for the 25 firms with the highest mean absolute distance to the average rating (MAD) within the normalized common sample (n=924). Firms are sorted by their average rating. Each rating agency is plotted in a different color.

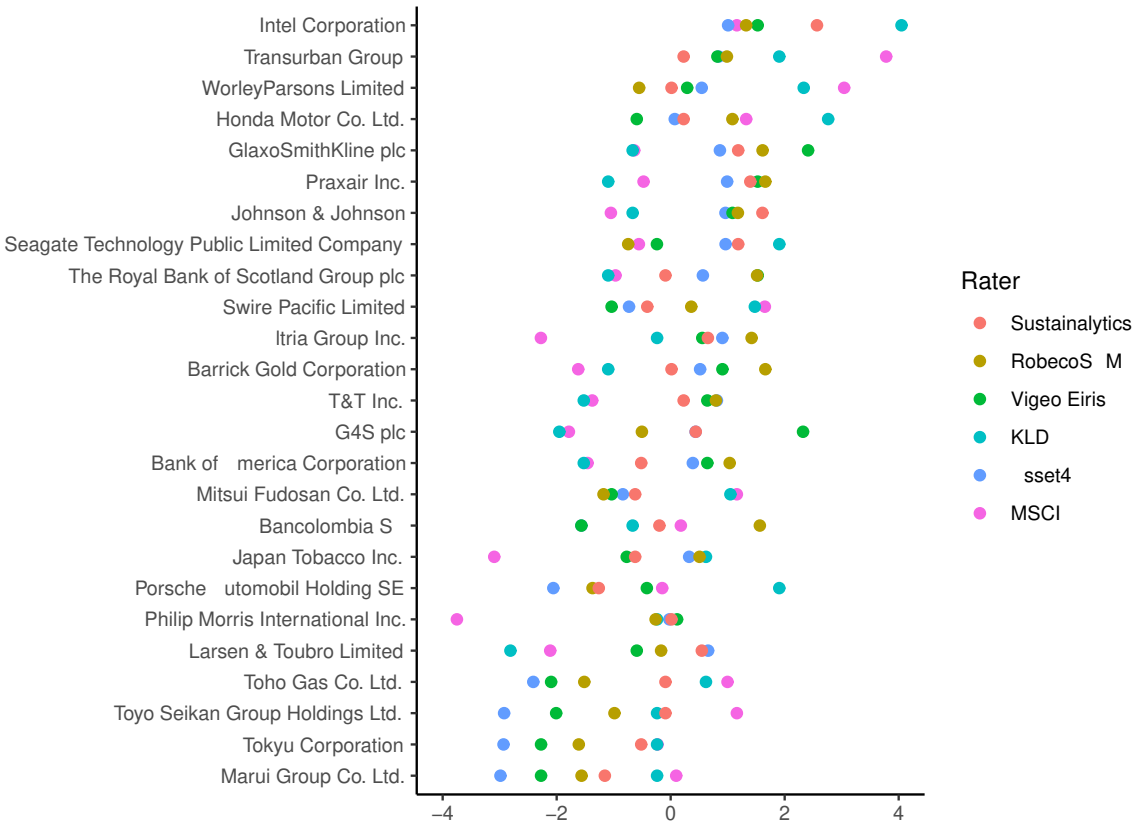


Figure 3
Quantile ranking count (QRC)

The orange line shows the QRC of the ESG ratings in the common sample and the dashed gray lines show the QRCs of simulated data. The QRC evaluates how many identical firms are included in the rating quantile across all six providers over the total number of firms. The size of the quantile is displayed on the x-axis and ranges from 5% to 100% in increments of 5%. The implied correlations are depicted by the gray lines, where the diagonal line reflects an implied correlation of 1 and the lowermost line reflects an implied correlation of 0.1. Implied correlations are shown for the values 1, 0.95, 0.9, and from then on in increments of 0.1.

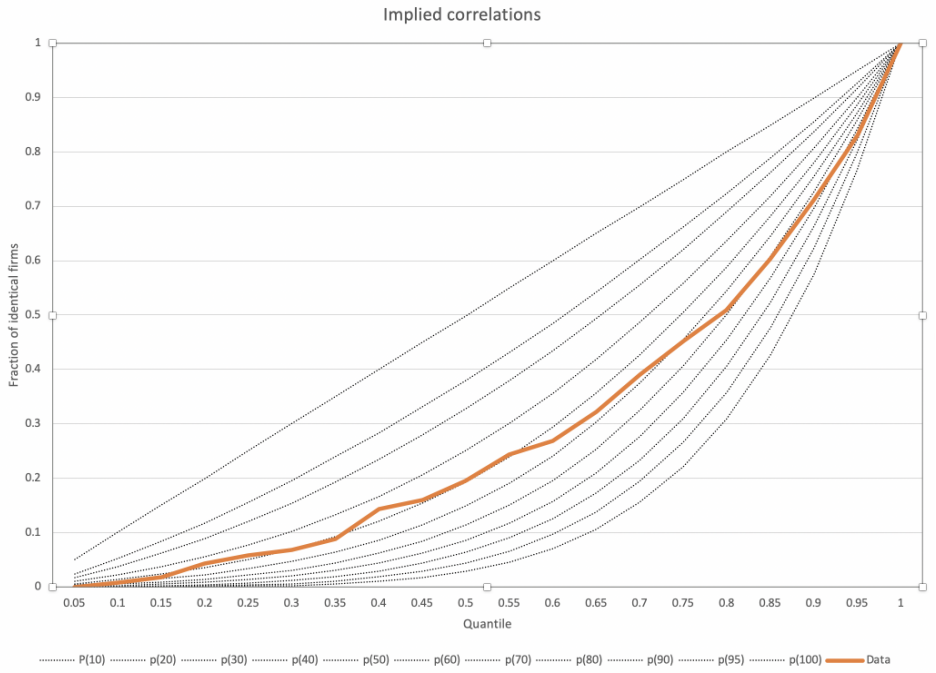


Figure 4
The sources of divergence

Our schematic representation of an ESG rating consists of the elements scope, measurement, and weights. Scope is the set of attributes A_n that describe a company's ESG performance. Measurement determines the indicators $I_{k\,1}$ ... $I_{k\,n}$, which produce numerical values for each attribute and are specific to rating agency k. Weights determine how indicators are aggregated into a single ESG rating R_k. Scope divergence results from two raters considering a different set of attributes. Measurement divergence results from two raters using different indicators to measure the same attribute. Weights divergence results from two raters aggregating the same indicators using different weights.

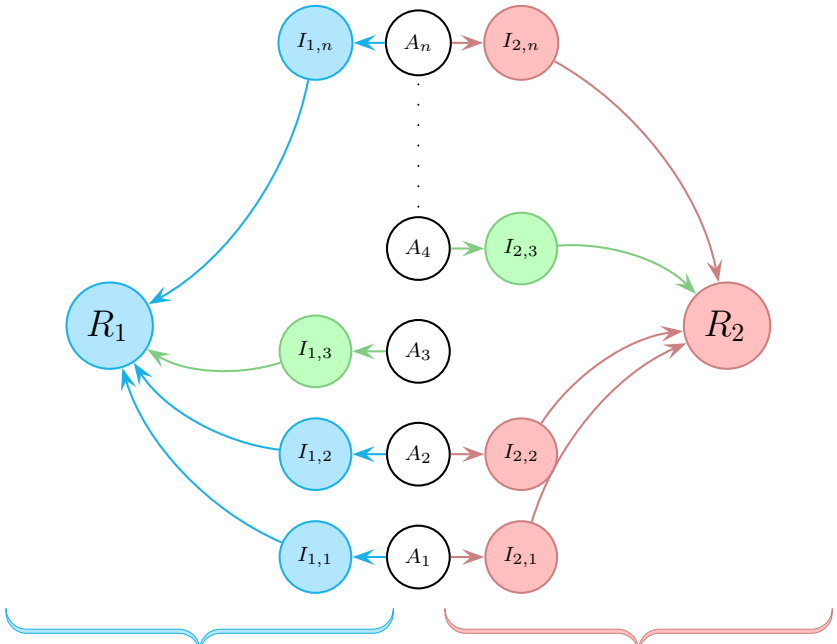


Figure 5
Decomposition example

Arithmetic decomposition of the difference between two ESG ratings, provided by Asset4 and KLD, for Barrick Gold Corporation in 2014. The normalized ratings are on the left and right. The overall divergence is separated into the contributions of scope divergence, measurement divergence, and weights divergence. Within each source the three most relevant categories in absolute terms are shown in descending order, with the remainder of the total value of each source labeled as "Other". The residual between the original rating and our fitted rating is shown in the second bar from the left and from the right, respectively.

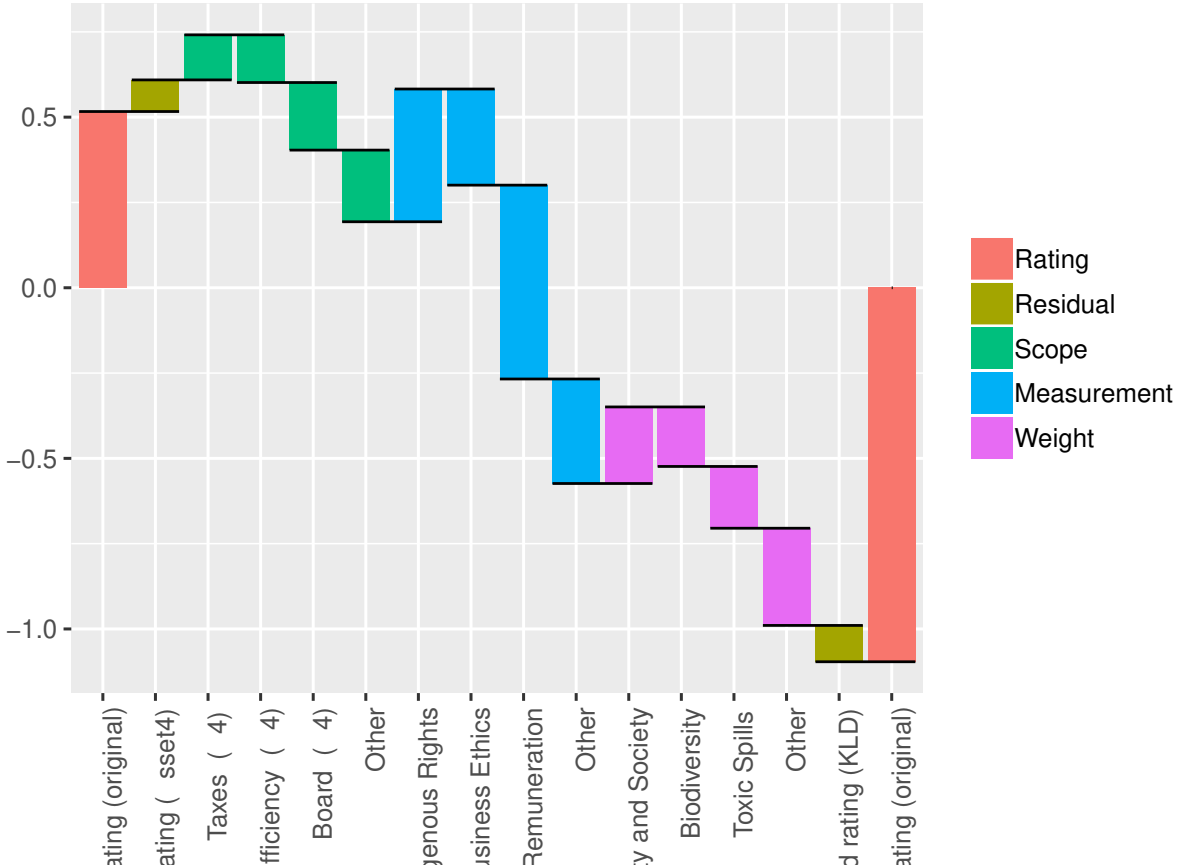
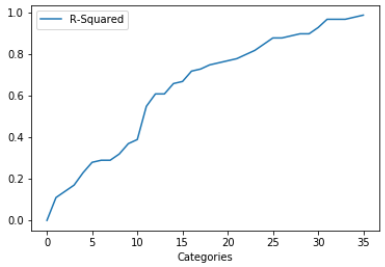
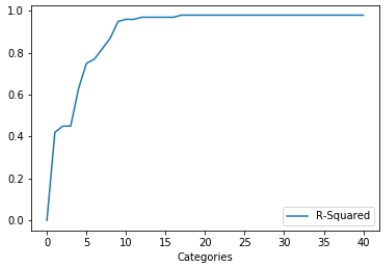


Figure 6
LASSO Regressions

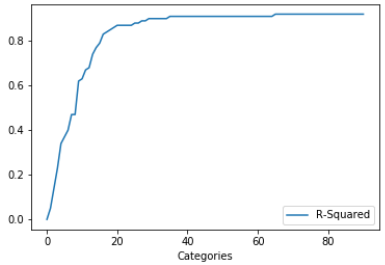
The plots show the R^2 values of a series of LASSO regressions, regressing the aggregate rating (ESG) of the different rating agencies on the categories of the same rater. The x-axis shows how many indicators are used as covariates and the y-axis indicates the corresponding R^2 value.



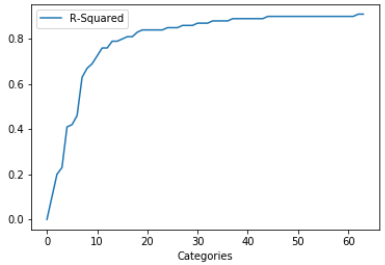
(a) KLD



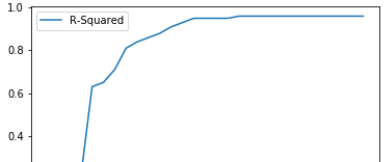
(b) RobecoSAM



(c) Asset4



(d) Sustainalytics



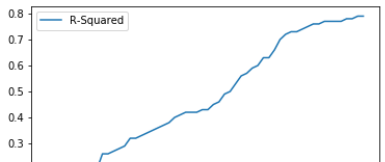


Table 1
Descriptive Statistics

Descriptive statistics of the aggregate rating (ESG) in 2014 for the six rating agencies. Panel A shows the data for the full sample, Panel B for the common sample, and Panel C for the normalized common sample.

	Sustainalytics	RobecoSAM	Vigeo Eiris	KLD	Asset4	MSCI
Panel A: Full Sample						
Firms	4531	1665	2304	5053	4013	9662
Mean	56.4	47.19	32.23	1.16	50.9	4.7
Standard Dev.	9.46	21.06	11.78	1.76	30.94	1.19
Minimum	29	13	5	-6	2.78	0
Median	55	40	31	1	53.15	4.7
Maximum	89	94	67	12	97.11	9.8
Panel B: Common Sample						
Firms	924	924	924	924	924	924
Mean	61.86	50.49	34.73	2.56	73.47	5.18
Standard Dev.	9.41	20.78	11.31	2.33	23.09	1.22
Minimum	37	14	6	-4	3.46	0.6
Median	62	47	33	2	81.48	5.1
Maximum	89	94	66	12	97.11	9.8
Panel C: Normalized Common Sample						
Firms	924	924	924	924	924	924
Mean	0.00	0.00	0.00	0.00	0.00	0.00
Standard Dev.	1.00	1.00	1.00	1.00	1.00	1.00
Minimum	-2.64	-1.76	-2.54	-2.81	-3.03	-3.75
Median	0.01	-0.17	-0.15	-0.24	0.35	-0.07
Maximum	2.89	2.09	2.76	4.05	1.02	3.78

Table 2
Correlations between ESG Ratings

Correlations between ESG ratings at the aggregate rating level (ESG) and at the level of the environmental dimension (E), the social dimension (S), and the governance dimension (G) using the common sample. The results are similar using pairwise common samples based on the full sample. SA, RS, VI, A4, KL, and MS are short for Sustainalytics, RobecoSAM, Vigeo Eiris, Asset4, KLD, and MSCI, respectively.

	KL SA	KL VI	KL RS	KL A4	KL MS	SA VI	SA RS	SA A4	SA MS	VI RS	VI A4	VI MS	RS A4	RS MS	A4 MS	Average
ESG	0.53	0.49	0.44	0.42	0.53	0.71	0.67	0.67	0.46	0.7	0.69	0.42	0.62	0.38	0.38	0.54
E	0.59	0.55	0.54	0.54	0.37	0.68	0.66	0.64	0.37	0.73	0.66	0.35	0.7	0.29	0.23	0.53
S	0.31	0.33	0.21	0.22	0.41	0.58	0.55	0.55	0.27	0.68	0.66	0.28	0.65	0.26	0.27	0.42
G	0.02	0.01	-0.01	-0.05	0.16	0.54	0.51	0.49	0.16	0.76	0.76	0.14	0.79	0.11	0.07	0.30

Table 3
Mean Absolute Distance (MAD)

This table shows how the mean absolute distance to the average rating (MAD) is distributed, based on the normalized common sample for the six rating agencies, KLD, Sustainalytics, Vigeo Eiris, RobecoSAM, Asset4, and MSCI. Panel A shows the distribution across regions and Panel B shows the distribution across industries. Since the ratings have been normalized to have zero mean and unit variance, all values can be interpreted in terms of standard deviations. Americas excludes Canada and USA. Asia excludes Japan.

Panel A: Distribution of Divergence

	MAD
Minimum	0.14
1st Quantile	0.36
Median	0.45
Mean	0.49
3rd Quantile	0.59
Maximum	1.26

Panel B: Average Divergence across Regions

	MAD	**No of Firms**
Africa	0.43	15
Americas	0.58	17
Asia	0.52	89
Canada	0.51	20
Europe	0.49	274
Japan	0.50	175
Oceania	0.47	53
USA	0.46	281

Panel C: Average Divergence across Industries

	MAD	**No of Firms**
Basic Materials	0.51	65
Consumer Discretionary	0.47	149
Consumer Staples	0.48	63
Energy	0.45	19
Financials	0.46	151

Table 4
Common Sets of Firms in Quantiles

The firms in this list are consistently included in the 20% (184) best- or worst-rated firms across all six rating agencies, using the common sample of 924 firms in 2014.

Common in Top Quantile	**Common in Bottom Quantile**
Akzo Nobel NV	Advance Auto Parts Inc.
Allianz SE	Affiliated Managers Group Inc.
Aviva plc	Amphenol Corporation
AXA Group	Anhui Conch Cement Co. Ltd.
Bayerische Motoren Werke Aktiengesellschaft	Cencosud SA
Dexus Property Group	China Development Financial Holding Corporation
Diageo plc	China Resources Land Ltd.
Industria de Diseno Textil SA	Credit Saison Co. Ltd.
Intel Corporation	Crown Castle International Corp.
Kingfisher plc	DR Horton Inc.
Koninklijke Philips NV	Expedia Inc.
SAP SE	Helmerich & Payne Inc.
Schneider Electric SA	Hengan International Group Company Limited
STMicroelectronics NV	Intuitive Surgical Inc.
Wipro Ltd.	Japan Real Estate Investment Corporation
	MediaTek Inc.
	NEXON Co. Ltd.
	Nippon Building Fund Inc.
	Ralph Lauren Corporation
	Shimano Inc.
	Sumitomo Realty & Development Co. Ltd.
	Sun Pharmaceutical Industries Limited
	Wynn Resorts Ltd.

Table 5
Number of Indicators per Rater and Category

This table shows how many indicators are provided by the different sustainability rating agencies per category.

	Sustainalytics	RobecoSAM	Asset4	Vigeo Eiris	MSCI	KLD
Access to Basic Services	2		1		1	1
Access to Healthcare	6	3	1		1	1
Animal Welfare	2		1			
Anti-competitive Practices			2	1	1	1
Audit	4		5	1		
Biodiversity	1	1	3	1	1	2
Board	6		25	1	1	
Board Diversity	2		1			3
Business Ethics	4	2	1		1	1
Chairperson-CEO Separation	1		1			
Child Labor			1	1		1
Climate Risk Mgmt.		2	1		1	2
Clinical Trials	1		1			
Collective Bargaining	2		1	1		
Community and Society	3	6	10	1		1
Corporate Governance		1			1	
Corruption	2		1	1	1	1
Customer Relationship	1	1	7	1		2
Diversity	2		9	1		3
ESG Incentives	1	1				
Electromagnetic Fields	1	1				
Employee Development	1	2	13	1	1	3
Employee Turnover	1		1			
Energy	3	6	5	1	2	1
Environmental Fines	1		1			1
Environmental Mgmt. System	2		1			1
Environmental Policy	4	2	4	2		
Environmental Reporting	2	1	1			
Financial Inclusion	1				1	1
Forests	1	1				
GHG Emissions	5		5	1		1
GHG Policies	3	2	4			
GMOs	1	1	1			
Global Compact Membership	1		1			
Green Buildings	5	2	1		1	1
Green Products	7	1	20	1	2	1
HIV Programs	1		1			
Hazardous Waste	1	1	1		1	
Health and Safety	7	1	7	1	1	2
Human Rights	2	1	5	1		5
Indigenous Rights	1		1			1
Labor Practices	3	1	16	4	1	3
Lobbying	3	1		1		
Non-GHG Air Emissions	1		2			
Ozone-Depleting Gases	1		1			
Packaging		1			1	1
Philanthropy	3	1	2	1		1
Privacy and IT	1	3			1	2
Product Safety	2	2	13	3	2	6
Public Health	1	3			1	2
Recycling					1	
Remuneration	4	1	15	2	1	4
Reporting Quality	3		5			1
Resource Efficiency	1	3	6			
Responsible Marketing	3	3	1	1		1
Shareholders			16	1		
Site Closure	1	1				

Table 6
Correlation of Category Scores

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 between different ratings agencies and categories being conditional on industries.

	KL SA	KL VI	KL RS	KL A4	KL MS	SA VI	SA RS	SA A4	SA MS	VI RS	VI A4	VI MS	RS A4	RS MS	A4 MS	Average
Access to Basic Services	0.08			0.13	0.85			0.49	0.15						0.16	0.31
Access to Healthcare	0.66		0.57	0.49	0.85		0.67	0.56	0.74				0.44	0.71	0.7	0.64
Animal Welfare								0.44								0.44
Anti-competitive Practices		-0.06		0.56	0.76						0	-0.05			0.56	0.30
Audit						0.57		0.66			0.62					0.62
Biodiversity		0.06	-0.08	0.06	0.66					0.61	0.41	0.47	0.47	0.01	0.2	0.29
Board						0.37		0.58			0.51					0.49
Board Diversity								0.8								0.80
Business Ethics	0.04		-0.11	0.4	0.6		0.33	0.03	0.01				-0.1	-0.15	0.38	0.14
Chairperson CEO Separation								0.59								0.59
Child Labor				0.49												0.49
Climate Risk Mgmt.			0.44	0.42	0.8								0.54	0.54	0.5	0.54
Clinical Trials								0.73								0.73
Collective Bargaining						0.59		-0.04			0					0.18
Community and Society	-0.15	0.25	0.2	0.11		-0.1	-0.19	-0.13		0.51	0.5		0.56			0.16
Corporate Governance														0.08		0.08
Corruption	0.26	0.24		-0.18	0.7	0.54		-0.19	0.37		-0.15	0.33			-0.12	0.18
Customer Relationship	0.38	-0.08	-0.09	0		-0.04	-0.13	-0.05		0.49	0.47		0.41			0.14
Diversity	-0.06	-0.02		0.03		0.61		0.52			0.56					0.27
ESG Incentives																
Electromagnetic Fields							0.68									0.68
Employee Development	0.22	0.29	0.37	0.37	0.73	0.23	0.19	0.36	0.34	0.39	0.29	0.31	0.55	0.45	0.51	0.37
Employee Turnover								0.4								0.40
Energy	0.22	0.13	0.49	0.25	0.8	0.4	0.27	0.27	0.4	0.32	0.41	0.59	0.2	0.4	0.48	0.38
Environmental Fines								0.05								0.05
Env. Mgmt. System	0.65			-0.09				0.46								0.34
Environmental Policy						0.52	0.46	0.46		0.63	0.61		0.62			0.55
Environmental Reporting							0.52	0.25					0.36			0.38
Financial Inclusion	0.29				0.7				0.51							0.50
Forests																
GHG Emissions	0	-0.03		-0.06		0.28		0.31			0.5					0.17
GHG Policies							0.48	0.62					0.41			0.50
GMOs							0.38	0.43					0.25			0.35
Global Compact Member								0.92								0.92
Green Buildings	0.54		0.59	0.21	0.83		0.25	0.26	0.55				-0.02	0.66	0.28	0.42
Green Products	0.23	0.07	0.27	0.34	0.76	0.1	0.37	0.47	0.32	0.31	0.29	-0.05	0.53	0.44	0.53	0.33
HIV Programs																
Hazardous Waste							0.22	0.13	0.34					0.59	0.1	0.28
Health and Safety	0.01	0.27	0.27	0.35	0.73	-0.1	-0.16	-0.16	-0.05	0.63	0.67	0.5	0.57	0.44	0.6	0.30
Human Rights	0	0.19		0.08		-0.01		-0.08			0.42					0.10
Indigenous Rights	0.26			-0.11				-0.46								-0.10
Labor Practices	0.21	-0.04	-0.14	0.07	0.1	0.2	0.14	0.32	0.27	0.54	0.45	0.43	0.35	0.34	0.37	0.24
Lobbying						-0.28										-0.28
Non-GHG Air Emissions								0.28								0.28
Ozone-Depleting Gases								0.44								0.44
Packaging																
Philanthropy						0.42	0.39	0.32		0.48	0.19		0.17			0.33
Privacy and IT	0.48		0.27		0.75		0.17		0.45					0.42		0.42
Product Safety	-0.05	0.06	0.16	0	0.63	-0.14		-0.03	0.07	0.46	0.21	0.11	0.38	-0.03	0.1	0.14
Public Health			0.6		0.74		0.38							0.63		0.59
Recycling																
Remuneration	0.15	0.09	-0.21	0.17		0.71	0.22	0.83		0.25	0.75		0.37			0.33
Reporting Quality								0.48								0.48
Resource Efficiency							0.35	0.42					0.57			0.45
Responsible Marketing	-0.5	-0.06	-0.38	0.24		0.38	0.68	0		0.49	0.05		-0.1			0.08

Table 7
Non-negative Least Squares Regression

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 before the normalization. The symbols ***,**, and * denote statistical significance at the 1, 5, and 10% levels, respectively. As the data was previously normalized, we exclude the constant term. The standard errors are bootstrapped. Non-existent categories are denoted by dashes.

	Sustainalytics	**RobecoSAM**	**Asset4**	**Vigeo Eiris**	**MSCI**	**KLD**
Access to Basic Services	0.019	-	0	-	0.138***	0.065***
Access to Healthcare	0.051***	0.004	0	-	0.079***	0.051***
Animal Welfare	0.05***	-	0	-	-	-
Anti - competitive Practices	-	-	0.05***	0.023***	0	0.131***
Audit	0	-	0.026*	0.084***	-	-
Biodiversity	0	0	0	0.028***	0.366***	0.076***
Board	0.072***	-	**0.196**	0.113***	0	-
Board Diversity	0.043***	-	0	-	-	0
Business Ethics	0.097***	0.046***	0.008	-	0	0.148***
Chairperson-CEO Separation	0.039***	-	0.016	-	-	-
Child Labor	-	-	0.008	0	-	0.046***
Climate Risk Mgmt.	-	**0.137**	0.064***	-	0.069**	**0.234**
Clinical Trials	0	-	0	-	-	-
Collective Bargaining	0.051***	-	0.011*	0.072***	-	-
Community and Society	0.079***	0.086***	0.03*	0.001	-	0.14***
Corporate Governance	-	0.048***	-	-	0.198***	-
Corruption	0.049***	-	0.022*	0.072***	**0.388**	0.124***
Customer Relationship	0.127***	0.097***	0.086***	0.027***	-	0.104***
Diversity	0.108***	-	0.066***	**0.159**	-	0.04***
ESG Incentives	0.006	0	-	-	-	-
Electromagnetic Fields	0.021**	0	-	-	-	-
Employee Development	0.018*	**0.221**	0.116***	0.067***	**0.406**	0.149***
Employee Turnover	0.024*	-	0	-	-	-
Energy	0.032**	0.016***	0.029**	0.103***	0.194***	0.046***
Environmental Fines	0	-	0	-	-	0
Environmental Mgmt. System	**0.199**	-	0.009	-	-	0.205***
Environmental Policy	0.091***	0.098***	0.012	**0.187**	-	-
Environmental Reporting	0.043**	0.039***	0.007	-	-	-
Financial Inclusion	0	-	-	-	0.089***	0.061***
Forests	0.008	0.016*	-	-	-	-
GHG Emissions	0.048***	-	0.002	0.033***	-	0.021**
GHG Policies	0.086***	0.008**	0.047**	-	-	-
GMOs	0	0	0	-	-	-
Global Compact Membership	0.029**	-	0	-	-	-
Green Buildings	0.072***	0.071***	0	-	0.304***	0.072***
Green Products	**0.167**	0.037***	0.093***	0.024**	0.351***	0.129***
HIV Programs	0	-	0.003	-	-	-
Hazardous Waste	0.021*	0	0	-	0.09***	-
Health and Safety	0.049***	0.042***	0.049***	**0.125**	0.148***	0.174***
Human Rights	0.072***	0	0.066***	0	-	0.14***
Indigenous Rights	0.033*	-	0.006	-	-	0.087***
Labor Practices	0.005	0.063***	0.067***	0.153***	0.166***	0.129***
Lobbying	0.091***	0	-	0.013	-	-
Non-GHG Air Emissions	0.014	-	0	-	-	-
Ozone-Depleting Gases	0	-	0	-	-	-
Packaging	-	0	-	-	0.128**	0.033***
Philanthropy	0.028*	0.075***	0.039***	0.073***	-	0
Privacy and IT	0.022*	0.039***	-	-	0.276***	0.124***
Product Safety	0.048***	0.002	0.059***	0.062***	**0.429**	**0.216**
Public Health	0.022**	0.011*	-	-	0.029	0.074***
Recycling	-	-	-	-	0.119***	-
Remuneration	0	0.054***	**0.117**	0.113***	0	**0.223**
Reporting Quality	0.123***	-	0.107***	-	-	0
Resource Efficiency	0.014	**0.114**	**0.135**	-	-	-
Responsible Marketing	0	0.033***	0	0.002	-	0.081***
Shareholders	-	-	0.111***	0.089***	-	-

Table 8
Quality of Fit

Comparison of the quality of fit in terms of R^2 for the estimation of rater-specific aggregation functions using different specifications. NNLS stands for non-negative least squares; OLS for ordinary least squares. NN stands for neural network with linear activation function, and NN Relu for a neural network with a non-linear relu activation function. RF stands for random forest. The symbol * indicates that the R^2 is reported for a testing set consisting of a randomly chosen 10% of the sample. The three last lines report results from the original method, but with different underlying data. For NNLS SASB the category scores were calculated based on the SASB taxonomy, for NNLS indicators the original indicators were used without any taxonomy, and for NNLS 2017 the underlying data is that of 2017 instead of that of 2014. Given that KLD does not offer any data for 2017, no value is reported.

Specification	KLD	Vigeo Eiris	RobecoSAM	Sustainalytics	MSCI	Asset4
NNLS	0.99	0.96	0.98	0.90	0.79	0.92
OLS	0.99	0.96	0.98	0.91	0.79	0.92
NNLS*	0.98	0.94	0.98	0.89	0.74	0.83
NN*	0.98	0.94	0.98	0.88	0.83	0.83
NN Relu*	0.96	0.96	0.98	0.83	0.85	0.80
RF*	0.73	0.91	0.97	0.85	0.56	0.86
NNLS SASB	0.98	0.96	0.98	0.87	0.76	0.92
NNLS Indicators	1	0.96	0.99	0.90	0.82	0.94
NNLS 2017		0.96	0.98	0.91	0.68	0.82

Table 9
Arithmetic Decomposition

Results from the arithmetic decomposition, which implements Equation 8, and relies on the category scores and estimated weights from Section 3. Panel A reports mean absolute values across firms for each pair of raters. The column "Scope" shows the difference between two ratings that were calculated on the basis of mutually exclusive categories. The column "Measurement" shows the difference between two ratings that were calculated on the basis of common categories and common weights. We estimate these common weights by jointly regressing the two ratings on the two raters' category scores. The column "Weights" shows the difference that results from exchanging the weights that were fitted for each rater individually with the common weights. The column "Fitted" shows the total difference between the fitted ratings, and "True" the total difference between the original ratings. For convenience, Panel B reports averages per rater on the basis of the values shown in Panel A.

Panel A: Rater Pairs

		Scope	Measurement	Weights	Fitted	True
KLD	Sustainalytics	0.27	0.6	0.29	0.73	0.76
KLD	Vigeo Eiris	0.4	0.6	0.27	0.78	0.79
KLD	RobecoSAM	0.28	0.67	0.31	0.8	0.81
KLD	Asset4	0.33	0.6	0.45	0.8	0.86
KLD	MSCI	0.85	0.51	0.51	0.71	0.77
Sustainalytics	Vigeo Eiris	0.39	0.51	0.24	0.54	0.6
Sustainalytics	RobecoSAM	0.32	0.55	0.16	0.58	0.64
Sustainalytics	Asset4	0.19	0.45	0.32	0.53	0.65
Sustainalytics	MSCI	0.86	0.52	0.53	0.76	0.82
Vigeo Eiris	RobecoSAM	0.3	0.39	0.11	0.6	0.61
Vigeo Eiris	Asset4	0.33	0.5	0.19	0.55	0.64
Vigeo Eiris	MSCI	0.78	0.55	0.43	0.81	0.85
RobecoSAM	Asset4	0.26	0.51	0.14	0.62	0.71
RobecoSAM	MSCI	0.86	0.6	0.57	0.83	0.89
Asset4	MSCI	0.85	0.57	0.56	0.78	0.89
Average		0.48	0.54	0.34	0.69	0.75

Panel B: Rater Averages

	Scope	Measurement	Weights	Fitted	True
KLD	0.43	0.60	0.37	0.76	0.80

Table 10
Range of Variance Explained

The first column presents the baseline R^2 for a regression of one rating onto another. The second column is the R^2 from a regression that includes all four covariates, i.e., it includes rating a plus the scope, measurement, and weight variables. The remaining columns indicate the minimum and maximum R^2 gain of explanatory power due the inclusion of the scope, measurement, and weight variables.

Panel A: Rater Pairs

			Scope		Measurement		Weights	
	Baseline	**All**	**Min**	**Max**	**Min**	**Max**	**Min**	**Max**
KLD on Sustainalytics	0 27	0 69	0 15	0 16	0 26	0 27	0 0	0 0
KLD on Vigeo Eiris	0 23	0 79	0 31	0 36	0 19	0 23	0 0	0 02
KLD on RobecoSAM	0 2	0 75	0 19	0 22	0 33	0 36	0 0	0 01
KLD on Asset4	0 21	0 75	0 02	0 05	0 44	0 5	0 01	0 06
KLD on MSCI	0 29	0 83	0 19	0 3	0 12	0 24	0 04	0 19
Sustainalytics on KLD	0 27	0 86	0 23	0 45	0 13	0 35	0 0	0 01
Sustainalytics on Vigeo Eiris	0 56	0 88	0 17	0 28	0 04	0 12	0 0	0 03
Sustainalytics on RobecoSAM	0 51	0 87	0 18	0 31	0 05	0 17	0 0	0 02
Sustainalytics on Asset4	0 55	0 8	0 01	0 03	0 13	0 18	0 06	0 11
Sustainalytics on MSCI	0 20	0 92	0 41	0 65	0 06	0 26	0 01	0 1
Vigeo Eiris on KLD	0 23	0 95	0 12	0 62	0 1	0 59	0 0	0 01
Vigeo Eiris on Sustainalytics	0 56	0 89	0 05	0 09	0 24	0 27	0 0	0 02
Vigeo Eiris on RobecoSAM	0 49	0 95	0 15	0 38	0 08	0 31	0 0	0 0
Vigeo Eiris on Asset4	0 56	0 9	0 0	0 04	0 3	0 34	0 0	0 01
Vigeo Eiris on MSCI	0 16	0 96	0 19	0 76	0 04	0 61	0 0	0 11
RobecoSAM on KLD	0 2	0 94	0 09	0 66	0 07	0 64	0 0	0 01
RobecoSAM on Sustainalytics	0 51	0 93	0 05	0 28	0 13	0 36	0 0	0 01
RobecoSAM on Vigeo Eiris	0 49	0 98	0 1	0 4	0 09	0 39	0 0	0 01
RobecoSAM on Asset4	0 48	0 95	0 01	0 05	0 36	0 46	0 0	0 07
RobecoSAM on MSCI	0 16	0 96	0 16	0 77	0 03	0 65	0 0	0 05
Asset4 on KLD	0 21	0 95	0 14	0 58	0 17	0 61	0 0	0 0
Asset4 on Sustainalytics	0 55	0 89	0 07	0 16	0 15	0 27	0 01	0 04
Asset4 on Vigeo Eiris	0 56	0 96	0 07	0 22	0 17	0 32	0 0	0 01
Asset4 on RobecoSAM	0 48	0 97	0 1	0 33	0 16	0 39	0 0	0 01
Asset4 on MSCI	0 18	0 89	0 18	0 69	0 01	0 51	0 0	0 12
MSCI on KLD	0 29	0 7	0 2	0 37	0 02	0 12	0 02	0 12
MSCI on Sustainalytics	0 2	0 44	0 15	0 22	0 02	0 08	0 0	0 01
MSCI on Vigeo Eiris	0 16	0 71	0 3	0 48	0 07	0 24	0 0	0 0
MSCI on RobecoSAM	0 16	0 63	0 13	0 37	0 1	0 35	0 0	0 01
MSCI on Asset4	0 18	0 57	0 17	0 31	0 08	0 22	0 0	0 02
Average	0 34	0 84	0 14	0 35	0 14	0 35	0 01	0 04

Panel B: Rater Averages

			Scope		Measurement		Weights	

Table 11
Investigation of Category and Rater E ect

The dependent variable is a vector that stacks all the common category scores for all raters, using the common sample. The independent variables are firm, firm-rater, and firm-category dummies. The difference in R^2 between regression 1 and 2 as well as 3 and 4 represents the rater effect.

Dummies	R^2
Firm	0.22
Firm + Firm–Rater	0.38
Firm + Firm–Category	0.47
Firm + Firm–Category + Firm–Rater	0.62

Table 12
LASSO Regressions

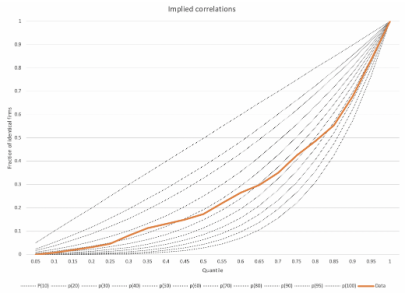
This table shows the R^2 of a series of LASSO regressions of ESG ratings on the categories of the same rater. The first column is the number of indicators that are used as covariates to obtain the corresponding R^2. The highlighted cells represent the number of categories that constitute 10% of the categories of the particular rating agency.

Categories	**Vigeo Eiris**	**RobecoSAM**	**Asset4**	**KLD**	**Sustainalytics**	**MSCI**
1	0.04	0.42	0	0.11	0.1	0.01
2	0.12	0.45	0.05	0.14	0.2	0.06
3	0.21	0.45	0.14	0.17	0.23	0.07
4	0.63	0.63	0.23	0.23	0.41	0.08
5	0.65	0.75	0.34	0.28	0.42	0.11
6	0.71	0.77	0.37	0.29	0.46	0.13
7	0.81	0.82	0.4	0.29	0.63	0.13
8	0.84	0.87	0.47	0.32	0.67	0.13

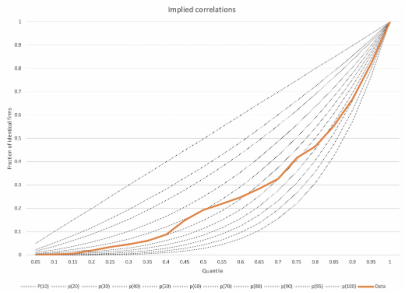
Internet Appendix

Figure A.1
Quantile Ranking Counts for Subdimensions

Analogous to Figure 3, but calculated separately for the Environmental, Social, and Governance Dimension of the ratings.



(a) Environment



(b) Social

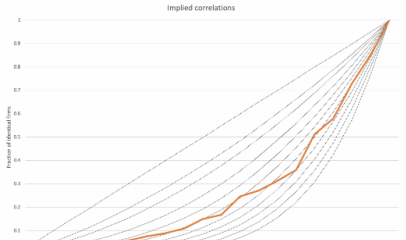


Table A.1
Number of Indicators per Rater and Category (SASB)

Calculation procedure and data are equivalent to Table 5, except that the underlying taxonomy is based on the 26 General Issue Categories provided by SASB.

	KLD	Sustainalytics	Vigeo Eiris	RobecoSAM	Asset4	MSCI
Access & Affordability	3	9		3	2	3
Air Quality		2			3	
Business Ethics	6	11	3	3	18	3
Business Model Resilience						
Competitive Behavior	1		1		2	1
Critical Incident Risk Mgmt.	1	1			2	
Customer Privacy	2	1		3		1
Customer Welfare	4	3	1	5	7	1
Data Security						
Ecological Impacts	3	11	3	6	9	1
Employee Engagement, Diversity & Inclusion	9	5	2	2	23	1
Employee Health & Safety	2	8	1	1	8	1
Energy Mgmt.	1	3	1	6	5	2
GHG Emissions	1	8	1	2	9	
Human Rights & Community Relations	7	6	2	7	16	
Labor Practices	5	6	6	1	20	1
Mgmt. of the Legal & Regulatory Env.	1	3		1	2	
Materials Sourcing & Efficiency		1		3	6	
Physical Impacts of Climate Change	2			2	1	1
Product Design & Lifecycle Mgmt.	3	16	1	6	23	4
Product Quality & Safety	6	2	3	2	13	2
Selling Practices & Product Labeling	1	3	1	3	1	
Supply Chain Mgmt.	6	21	4	3	4	3
Systemic Risk Mgmt.	1			1		1
Waste & Hazardous Materials Mgmt.	3	4	1	3	5	1
Water & Wastewater Mgmt.	2	2	1	2	3	1
Unclassified	8	37	6	15	100	40
Sum	78	163	38	80	282	68

Table A.2
Correlation of Category Scores (SASB)

Calculation procedure and data are equivalent to Table 6, except that the underlying taxonomy is based on the 26 General Issue Categories provided by SASB.

	KL SA	KL VI	KL RS	KL A4	KL MS	SA VI	SA RS	SA A4	SA MS	VI RS	VI A4	VI MS	RS A4	RS MS	A4 MS	Average
Access & Affordability	0.41		0.57	0.25	0.78		0.67	0.47	0.56				0.55	0.71	0.43	0.54
Air Quality								0.27								0.27
Business Ethics	0.1	0.05	0.05	0.09	0.35	0.68	0.43	0.73	0.08	0.43	0.69	0.05	0.25	0.17	0.01	0.28
Competitive Behavior		-0.06		0.56	0.76						0	-0.05			0.56	0.30
Critical Incident Risk Mgmt.				0.21												0.21
Customer Privacy	0.48		0.27		0.75		0.17		0.45					0.42		0.42
Customer Welfare	0.31	-0.08	-0.06	0.02	0.67	-0.03	-0.11	-0.04	-0.07	0.48	0.47		0.42	0.4	0.38	0.20
Ecological Impacts	0.48	0.4	0.41	0.39	0.65	0.67	0.7	0.65	0.29	0.71	0.58	0.48	0.69	0.21	0.26	0.50
Employee Engagement, Diversity & Inclusion	0.17	0.2	0.15	0.2	0.72	0.57	0.4	0.54	0.45	0.51	0.55	0.42	0.58	0.45	0.55	0.43
Employee Health & Safety	0.01	0.27	0.27	0.34	0.73	-0.11	-0.16	-0.14	-0.06	0.63	0.66	0.5	0.55	0.44	0.59	0.30
Energy Mgmt.	0.22	0.13	0.49	0.25	0.8	0.4	0.27	0.27	0.4	0.32	0.41	0.59	0.2	0.4	0.48	0.38
GHG Emissions	0	-0.03		-0.06		0.32	0.63	0.59			0.56		0.36			0.30
Human Rights & Community Relations	-0.13	0.25	0.15	0.11		-0.03	-0.14	-0.09		0.54	0.49		0.64			0.18
Labor Practices	0.26	0.28	0.11	0.2	0.34	0.59	0.45	0.42	0.41	0.56	0.48	0.43	0.38	0.34	0.4	0.38
Mgmt. of the Legal & Regulatory Environment							0.05	-0.05					0.01			0.00
Materials Sourcing & Efficiency							0.35	0.42					0.57			0.45
Physical Impacts of Climate Change			0.44	0.42	0.8								0.54	0.54	0.5	0.54
Product Design & Lifecycle Mgmt.	0.29	0.07	0.31	0.29	0.78	0.31	0.47	0.35	0.42	0.35	0.3	-0.05	0.56	0.48	0.48	0.36
Product Quality & Safety	-0.05	0.06	0.16	0	0.63	-0.14		-0.03	0.07	0.46	0.21	0.11	0.38	-0.03	0.1	0.14
Selling Practices & Product Labeling	-0.5	-0.06	-0.38	0.24		0.38	0.68	0		0.49	0.05		-0.1			0.08
Supply Chain Mgmt.	0.15	0.17	0.13	0.16	0.62	0.57	0.53	0.56	0.61	0.66	0.62	0.6	0.53	0.34	0.48	0.45
Systemic Risk Mgmt.			0.24		0.65									0.24		0.38
Waste & Hazardous Materials Mgmt.	0.25	0.34	0.22	0.23	0.78	0.43	0.22	0.36	0.33	0.48	0.32	0.39	0.12	0.23	0.3	0.33
Water & Wastewater Mgmt.	0.36	0.36	0.23	0.23	0.67	0.47	0.29	0.31	0.45	0.48	0.32	0.5	-0.02	0.24	0.44	0.36
Average	0.17	0.15	0.21	0.22	0.68	0.34	0.33	0.29	0.31	0.51	0.42	0.33	0.38	0.35	0.40	

Table A.3
Non Negative Least Squares Regression (SASB)

Calculation procedure and data is equivalent to Table 7, except that the underlying taxonomy is based on the 26 General Issue Categories provided by SASB.

	Sustainalytics	**RobecoSAM**	**Asset4**	**Vigeo Eiris**	**MSCI**	**KLD**
Access & Affordability	0.032**	0	0	-	0.207***	0.099***
Air Quality	0.022*	-	0	-	-	-
Business Ethics	0.12***	0.059***	0.098***	0.186***	0.055*	0.273***
Competitive Behavior	-	-	0.049***	0.01	0	0.134***
Critical Incident Risk Mgmt.	0	-	0	-	-	0.106***
Customer Privacy	0.033***	0.04***	-	-	0.27***	0.122***
Customer Welfare	0.131***	0.072***	0.089***	0.031***	0.031	0.118***
Ecological Impacts	0.322***	0.156***	0.007	0.19***	0.419***	0.216***
Employee Engagement, Diversity & Inclusion	0.08***	0.226***	0.152***	0.198***	0.406***	0.139***
Employee Health & Safety	0.019	0.056***	0.051***	0.133***	0.174***	0.178***
Energy Mgmt.	0.037***	0.004	0.028*	0.101***	0.211***	0.054***
GHG Emissions	0.144***	0.01***	0.03	0.036***	-	0.024***
Human Rights & Community Relations	0.101***	0.084***	0.079***	0.03***	-	0.31***
Labor Practices	0.075***	0.064***	0.072***	0.189***	0.149***	0.209***
Mgmt. of the Legal & Regulatory Environment	0.023*	0.004	0.005	-	-	0
Materials Sourcing & Efficiency	0.013	0.095***	0.133***	-	-	-
Physical Impacts of Climate Change	-	0.14***	0.069***	-	0.089***	0.238***
Product Design & Lifecycle Mgmt.	0.05***	0.052***	0.101***	0.01	0.484***	0.138***
Product Quality & Safety	0.065***	0	0.064***	0.064***	0.427***	0.219***
Selling Practices & Product Labeling	0	0.031***	0	0	-	0.086***
Supply Chain Mgmt.	0.245***	0.053***	0.049***	0.037***	0.163***	0.122***
Systemic Risk Mgmt.	-	0.059***	-	-	0.362***	0.106***
Waste & Hazardous Materials Mgmt.	0.059***	0.016*	0.032**	0.001	0.077**	0.193***
Water & Wastewater Mgmt.	0.066***	0.017**	0.029**	0	0.039*	0.176***
Unclassified Indicators	Yes	Yes	Yes	Yes	Yes	Yes
R2	0.87	0.98	0.92	0.96	0.76	0.98
Observations	924	924	924	924	924	924

Table A.4
Arithmetic Decomposition (SASB)

Calculation procedure and data equivalent to Table 9, except that the underlying taxonomy is based on the 26 General Issue Categories provided by SASB.

		Scope	Measurement	Weights	Residuals	Fitted	True
KLD	**Sustainalytics**	0 25	0 66	0 29	0 3	0 72	0 76
KLD	**Vigeo Eiris**	0 33	0 72	0 31	0 18	0 79	0 79
KLD	**RobecoSAM**	0 23	0 71	0 42	0 15	0 81	0 81
KLD	**Asset**	0 36	0 6	0 48	0 25	0 8	0 86
KLD	**MSCI**	0 8	0 58	0 45	0 4	0 71	0 77
Sustainalytics	**Vigeo Eiris**	0 21	0 47	0 19	0 32	0 53	0 6
Sustainalytics	**RobecoSAM**	0 18	0 48	0 21	0 3	0 57	0 64
Sustainalytics	**Asset**	0 29	0 39	0 25	0 36	0 52	0 65
Sustainalytics	**MSCI**	0 75	0 62	0 44	0 48	0 76	0 82
Vigeo Eiris	**RobecoSAM**	0 28	0 45	0 16	0 18	0 59	0 61
Vigeo Eiris	**Asset**	0 31	0 41	0 2	0 28	0 55	0 64
Vigeo Eiris	**MSCI**	0 72	0 72	0 41	0 4	0 81	0 85
RobecoSAM	**Asset**	0 26	0 47	0 1	0 26	0 62	0 71
RobecoSAM	**MSCI**	0 8	0 68	0 52	0 39	0 83	0 89
Asset	**MSCI**	0 84	0 55	0 54	0 46	0 79	0 89
Average		0 44	0 57	0 33	0 31	0 69	0 75

Table A.5
Range of Variance Explained (SASB)

Calculation procedure and data equivalent to Table 10, except that the underlying taxonomy is based on the 26 General Issue Categories provided by SASB.

	Baseline	All Covariates	Scope		Measurement		Weights	
			Min	Max	Min	Max	Min	Max
KLD on Sustainalytics	28	76	1	12	36	38		1
KLD on Vigeo Eiris	22	76	9	14	37	42	1	4
KLD on RobecoSAM	19	75	2	5	51	53		1
KLD on Asset4	2	8	3	5	53	56	1	4
KLD on MSCI	29	79	11	12	21	34	5	18
Sustainalytics on KLD	28	86	1	32	26	48		1
Sustainalytics on Vigeo Eiris	56	87	3	1	17	25	2	4
Sustainalytics on RobecoSAM	51	89	4	14	22	33	1	2
Sustainalytics on Asset4	56	86	5	14	12	2	2	5
Sustainalytics on MSCI	19	89	13	53	15	56		9
Vigeo Eiris on KLD	22	95	9	31	41	64		
Vigeo Eiris on Sustainalytics	56	95	8	14	24	28		3
Vigeo Eiris on RobecoSAM	5	97	8	19	28	39		
Vigeo Eiris on Asset4	56	95	5	8	26	34		5
Vigeo Eiris on MSCI	14	94	13	42	36	66		7
RobecoSAM on KLD	19	97	4	65	14	7		2
RobecoSAM on Sustainalytics	51	97	9	31	15	37		1
RobecoSAM on Vigeo Eiris	5	98	4	38	1	44		3
RobecoSAM on Asset4	48	99	3	15	26	48		1
RobecoSAM on MSCI	14	96	1	76	5	71		6
Asset4 on KLD	2	98	12	63	15	66		
Asset4 on Sustainalytics	56	96	14	28	13	26		2
Asset4 on Vigeo Eiris	56	97	1	33	9	32		
Asset4 on RobecoSAM	48	98	12	32	18	38		1
Asset4 on MSCI	16	97	23	74	7	58		8
MSCI on KLD	29	67	21	34	2	11	2	6
MSCI on Sustainalytics	19	52	17	3	2	14		1
MSCI on Vigeo Eiris	14	54	18	31	8	21		1
MSCI on RobecoSAM	14	62	14	36	11	33		1
MSCI on Asset4	16	69	19	43	1	34		1
Average	33	86	1	3	2	41		4

Table A.6
Correlations between ESG Ratings (2017)

Calculation procedure and data equivalent to Table 10, except that the underlying data is from 2017 instead of from 2014.

	SA	SA	SA	SA	VI	VI	VI	RS	RS	A4	Average

Table A.7
Correlation of Category Scores (2017)

Calculation procedure and data equivalent to Table 6, except that the underlying data is from 2017 instead of from 2014.

	SA VI	SA RS	SA A4	SA MS	VI RS	VI A4	VI MS	RS A4	RS MS	A4 MS	Average
Access to Basic Services			0.44	-0.08						0.23	0.20
Access to Healthcare		0.58	0.73	0.67				0.4	0.55	0.7	0.61
Animal Welfare			0.62								0.62
Anti-competitive Practices						0.01	0.08			0.44	0.18
Audit	0.46		0.66			0.41					0.51
Biodiversity	0.61	0.7	0.34	0.41	0.55	0.35	0.4	0.36	0.33	0.27	0.43
Board	0.35		0.61	0.36		0.45	0.43			0.34	0.42
Board Diversity			0.75								0.75
Business Ethics		0.31	0.06	0.2				-0.05	-0.04	0.39	0.15
Chairperson CEO Separation			0.59								0.59
Child Labor											
Climate Risk Mgmt.								0.42	0.5	0.32	0.41
Clinical Trials			0.5								0.50
Collective Bargaining	0.62		-0.05			0					0.19
Community and Society	-0.06	-0.14	-0.07		0.5	0.43		0.52			0.20
Corporate Governance									0.39		0.39
Corruption	0.53		-0.22	0.39		-0.1	0.47			-0.07	0.17
Customer Relationship	-0.07	-0.09	-0.06		0.49	0.43		0.42			0.19
Diversity	0.66		0.56			0.56					0.59
ESG Incentives		0.48									0.48
Electromagnetic Fields		0.41									0.41
Employee Development	-0.15	0.29	0.34	0.29	0.32	0.26	0.17	0.49	0.37	0.42	0.28
Employee Turnover			0.46								0.46
Energy	0.4	0.22	0.26	0.37	0.19	0.33	0.05	0.02	0.17	0.36	0.24
Environmental Fines			0.28								0.28
Environmental Mgmt. System			0.5								0.50
Environmental Policy	0.53	0.5	0.46		0.6	0.54		0.54			0.53
Environmental Reporting		0.52	0.25					0.37			0.38
Financial Inclusion				0.43							0.43
Forests											
GHG Emissions	0.25		0.28			0.47					0.33
GHG Policies		0.31	0.64					0.2			0.38
GMOs		0.46	0.61					0.01			0.36
Global Compact Membership			0.83								0.83
Green Buildings		0.22	0.19	0.55				0.18	0.34	0.34	0.30
Green Products	0.46	0.42	0.53	0.34	0.22	0.31	0.22	0.46	0.36	0.5	0.38
HIV Programs			0.75								0.75
Hazardous Waste		0.16	0.05	0.37					0.27	0.1	0.19
Health and Safety	-0.06	-0.07	-0.13	0	0.54	0.66	0.55	0.49	0.39	0.6	0.30
Human Rights	0.02	0.01	-0.04		0.43	0.41		0.45			0.21
Indigenous Rights			-0.22								-0.22
Labor Practices	0.43	0.28	0.24	0.16	0.53	0.35	0.3	0.24	0.19	0.36	0.31
Lobbying	-0.28	-0.34			0.41						-0.07
Non-GHG Air Emissions			0.45								0.45
Ozone-Depleting Gases			0.41								0.41
Packaging											
Philanthropy					0.42	0.27		0.16			0.28
Privacy and IT		0.16		0.33					0.31		0.27
Product Safety	-0.11	-0.12	-0.01	0.04	0.27	0.19	0.23	0.21	0.33	0.31	0.13
Public Health		0.55		0.13					0.26		0.31
Recycling											
Remuneration	0.65	-0.03	0.8	0.2	0.2	0.7	0.27	-0.02	0.1	0.26	0.31
Reporting Quality			0.47								0.47
Resource Efficiency		0.2	0.27					0.5			0.32

Table A.8
Non Negative Least Squares Regression (2017)

Calculation procedure and data equivalent to Table 7, except that the underlying data is from 2017 instead of from 2014.

	Sustainalytics	RobecoSAM	Asset4	Vigeo Eiris	MSCI
Access to Basic Services	0.026**	-	0	-	0.117***
Access to Healthcare	0.062***	0.013**	0	-	0.08***
Animal Welfare	0.034***	-	0	-	-
Anti-competitive Practices	-	-	0.037	0.019**	0
Audit	0	-	0	0.062***	-
Biodiversity	0	0	0	0.019***	0.244***
Board	0.093***	-	0.21***	0.112***	0.028
Board Diversity	0	-	0.02	-	-
Business Ethics	0.104***	0.097***	0	-	0
Chairperson-CEO Separation	0.048***	-	0	-	-
Child Labor	-	-	0	0	-
Climate Risk Mgmt.	-	0.151***	0.012	-	0.146***
Clinical Trials	0	-	0.006	-	-
Collective Bargaining	0.081***	-	0	0.068***	-
Community and Society	0.072***	0.057***	0.029	0.014**	-
Corporate Governance	-	0.037***	-	-	0.265***
Corruption	0.029***	-	0.039**	0.088***	0.476***
Customer Relationship	0.093***	0.044***	0.059***	0.03***	-
Diversity	0.087***	-	0.027	0.126***	-
ESG Incentives	0.01	0.061***	-	-	-
Electromagnetic Fields	0.004	0	-	-	-
Employee Development	0	0.193***	0.118***	0.062***	0.437***
Employee Turnover	0.044***	-	0.043***	-	-
Energy	0.028***	0.021***	0.062***	0.133***	0.194***
Environmental Fines	0	-	0	-	-
Environmental Mgmt. System	0.194***	-	0	-	-
Environmental Policy	0.071***	0.069***	0.029*	0.18***	-
Environmental Reporting	0.04***	0.058***	0.003	-	-
Financial Inclusion	0	-	-	-	0.086***
Forests	0	0.006*	-	-	-
GHG Emissions	0.044***	-	0	0.042***	-
GHG Policies	0.086***	0	0	-	-
GMOs	0	0	0	-	-
Global Compact Membership	0.044***	-	0	-	-
Green Buildings	0.089***	0.039***	0.006	-	0.169***
Green Products	0.158***	0.017***	0.049***	0.055***	0.227***
HIV Programs	0	-	0	-	-
Hazardous Waste	0.013	0	0	-	0.016
Health and Safety	0.094***	0.008	0.016	0.108***	0.104***
Human Rights	0.017**	0.039***	0.048***	0.018**	-
Indigenous Rights	0.03**	-	0	-	-
Labor Practices	0.019**	0.03***	0.023	0.147***	0.131***
Lobbying	0.093***	0.03***	-	0.005	-
Non-GHG Air Emissions	0.011	-	0.006	-	-
Ozone-Depleting Gases	0	-	0	-	-
Packaging	-	0	-	-	0.14***
Philanthropy	0	0.069***	0.101***	0.068***	-
Privacy and IT	0.018*	0.026***	-	-	0.356***
Product Safety	0.047***	0	0.039	0.025***	0.094***
Public Health	0.005	0	-	-	0
Recycling	-	-	-	-	0.07***
Remuneration	0	0.026***	0.129***	0.101***	0
Reporting Quality	0.134***	-	0.1***	-	-
Resource Efficiency	0.003	0.114***	0.137***	-	-
Responsible Marketing	0	0.025***	0	0	-
Shareholders	-	-	0.119***	0.084***	-
Site Closure	0	0.031***	-	-	-

Table A.9
Arithmetic Decomposition (2017)

Calculation procedure and data equivalent to Table 9, except that the underlying data is from 2017 instead of from 2014.

		Scope	Measurement	Weights	Residuals	Fitted	True
Sustainalytics	**Vigeo Eiris**	0 42	0 5	0 19	0 26	0 49	0 53
Sustainalytics	**RobecoSAM**	0 32	0 53	0 16	0 24	0 6	0 65
Sustainalytics	**Asset**	0 19	0 45	0 35	0 36	0 64	0 76
Sustainalytics	**MSCI**	0 8	0 44	0 47	0 48	0 7	0 76
Vigeo Eiris	**RobecoSAM**	0 39	0 38	0 17	0 18	0 66	0 68
Vigeo Eiris	**Asset**	0 28	0 54	0 21	0 32	0 6	0 71
Vigeo Eiris	**MSCI**	0 69	0 5	0 35	0 44	0 72	0 8
RobecoSAM	**Asset**	0 3	0 56	0 15	0 29	0 79	0 9
RobecoSAM	**MSCI**	0 83	0 6	0 54	0 43	0 82	0 88
Asset	**MSCI**	0 77	0 45	0 47	0 51	0 71	0 86
Average		0 5	0 5	0 31	0 35	0 67	0 75

Table A.10
Range of Variance Explained (2017)

Calculation procedure and data equivalent to Table 10, except that the underlying data is from 2017 instead of from 2014.

			Scope		Measurement		Weights	
	Baseline	All Covariates	Min	Max	Min	Max	Min	Max
Sustainalytics on Vigeo Eiris	0.62	0.92	0.2	0.23	0.07	0.09	0.0	0.01
Sustainalytics on RobecoSAM	0.47	0.88	0.22	0.38	0.04	0.19	0.0	0.01
Sustainalytics on Asset4	0.36	0.69	0.02	0.05	0.12	0.16	0.16	0.18
Sustainalytics on MSCI	0.28	0.93	0.39	0.58	0.05	0.24	0.01	0.03
Vigeo Eiris on Sustainalytics	0.62	0.9	0.03	0.07	0.20	0.25	0.0	0.02
Vigeo Eiris on RobecoSAM	0.4	0.95	0.17	0.46	0.09	0.37	0.0	0.0
Vigeo Eiris on Asset4	0.42	0.88	0.0	0.04	0.4	0.45	0.0	0.02
Vigeo Eiris on MSCI	0.26	0.97	0.13	0.65	0.05	0.57	0.0	0.09
RobecoSAM on Sustainalytics	0.47	0.94	0.07	0.39	0.07	0.4	0.0	0.0
RobecoSAM on Vigeo Eiris	0.40	0.98	0.15	0.55	0.03	0.43	0.0	0.01
RobecoSAM on Asset4	0.23	0.94	0.02	0.52	0.14	0.69	0.0	0.20
RobecoSAM on MSCI	0.17	0.98	0.14	0.77	0.04	0.67	0.0	0.03
Asset4 on Sustainalytics	0.36	0.86	0.17	0.37	0.12	0.31	0.0	0.05
Asset4 on Vigeo Eiris	0.42	0.94	0.04	0.35	0.17	0.47	0.0	0.04