Chasing ESG Performance: Revealing the Impact of Refinitiv's Scoring System

Matteo Benuzzi[1,*], Karoline Bax[2,*], Sandra Paterlini[1], and Emanuele Taufer[1]

[1]Department of Economics and Management, University of Trento, Via Inama 5, 38122 Trento, Italy
[2]Center for Digital Transformation, Technical University of Munich, Bildungscampus 9, 74076 Heilbronn, Germany
[*]Equal Contribution

December 12, 2023

Abstract

ESG metrics, central to sustainable investing, have witnessed heightened scrutiny for accuracy and representation. Recently, concerns arise from potential retroactive score adjustments and data aggregation that might misrepresent company performances. Our study spans three key sectors from 2017-2021, utilizing a comprehensive set of ESG data. We reveal that Refinitiv's current rating methodology might obscure genuine company progress by a combination of artificially inflating high-ranking firms' scores with the introduction of new less-performing firms, and at the same time diminishing their actual advancements due to the competition among peers. To address these issues, we propose to replace the percentile ranking methodology within Refinitiv approach with the so-called 'performance ratio' scoring methodology. Our analyses shows a significant correlation between Refinitiv's existing approach and our performance ratio method, yet the latter is much less affected by new entrants and peer comparison, while maintaining an overall robustness towards outliers. Our findings highlight the need for a refined ESG scoring system that more accurately mirrors corporate sustainability efforts and actual underlying data.

1 Introduction

Environmental, Social, and Governance (ESG) metrics have become a focal point in contemporary discussions about sustainable and responsible investing. With the looming challenges posed by climate change, the appeal of ESG frameworks as easy-to-use tools for investors becomes evident. They offer a structured way to assess a company's commitment to ethical practices, environmental preservation, and good governance. While the market is now saturated with numerous ESG data providers, Refinitiv (2023)—recently rebranded as LSEG following its acquisition two years ago—remains one of the most frequently utilized. Intriguingly, Refinitiv distinguishes itself by granting its subscribers complete access to all data points and the associated questions used in their ESG score derivation.

Refinitiv's ESG scoring model is structured around three core pillars: Environmental (E), Social (S), and Governance (G), which are further divided into ten subpillars. The Environmental and Governance pillars are composed of three subpillars each, whereas the Social pillar encompasses four (see Figure 3). Importantly, the weighting of these pillars and subpillars varies by sector, a detail that Refinitiv openly discloses.

Recent studies, such as Berg, Fabisik, and Sautner (2020), have raised questions about the accuracy of ESG metrics, particularly concerning retroactive score adjustments. Given the vast amount of data collected, significant portions that remain missing (see e.g. Sahin, Bax, Czado, and Paterlini (2022)) or are complexly aggregated can result in distorted representations. A key issue is the double application of percentile ranking in calculating subpillar scores in the methodology of Refinitiv, which can obscure performance differences between companies and exaggerate score variances, thereby diminishing nuanced distinctions.

This issue becomes noticeable when we plot the yearly ESG scores of a large set of global companies per sector and compare these plots in Figure 1. The figure is organized into six boxplots, arranged across three columns and two rows, each representing data from 2017 to 2021. The columns categorize the companies into three sectors: (i) Machinery, Tools, Heavy Vehicles, Trains, and Ships, (ii) Oil & Gas, and (iii) Chemicals. In the top row, the plot presents a broad universe of global companies, including a company in the boxplot for a particular year only if it has an ESG score for that year. This results in varying company representations across different years. Contrastingly, the second row adopts a stricter criterion, showcasing only those companies that have sustained a continuous ESG score from 2017 to 2021. It is clear that companies that have been in the sample longer (row 2), have improved their scores over time, suggesting the capability to reach higher level of sustainability. However, we question if this improvement is due to the entrance of new companies in the sample joint to the use of percentile ranking schemes in building the final score or to an actual improvement in the variables that enter the ESG score computation or to the joint effect.

This paper critically examines the validity and reliability of Refinitiv's percentile ranking aggregations in ESG scoring, which was introduced to better deal with potential outliers. However, our analysis suggests that the current Refinitiv method may not only obscure actual company progress but also potentially create an illusion of advancement where none exists. This could occur through a dual mechanism: by artificially inflating the scores of high-ranking companies,

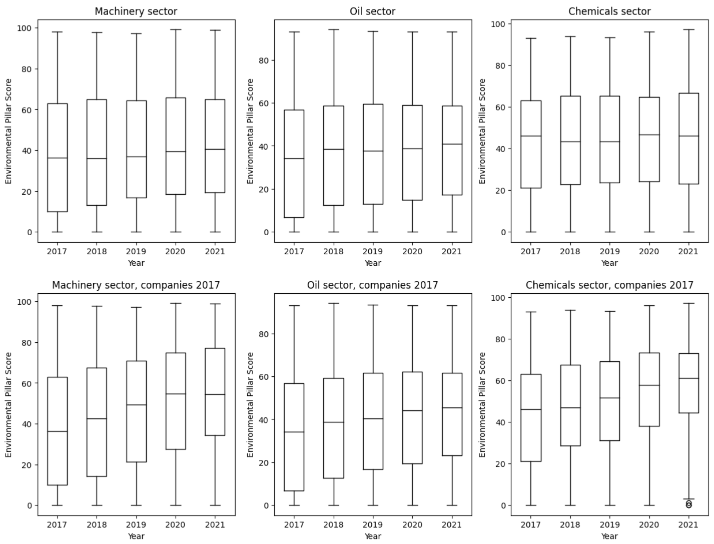


Figure 1: Boxplots of the Environmental Pillar Scores in the Machinery (column 1), Oil (column 2) and Chemicals (column 3) sectors analyzed in the 2017-2021. The first row refers to the full sample in 2021, while the second row refers to the initial sample (companies that were in the sample in 2017). The increasing trend of the Environmental scores is clearly visible in the second row. Similar findings for individual ESG pillars.

where new entrants—often lacking sufficient disclosure information—can distort the representation of ESG performance for leaders, and concurrently downplaying their advancements relative to their peer development. We observe that this phenomenon is facilitated by the methodology's dual percentile ranking. Such practices could create a misleading impression of a company's ESG performance progress. When analyzing a company's time series data and noticing an improvement over time, investors might not be able to discern if any real improvement has occurred in the company's disclosed ESG information without a detailed analysis of the single indicators. This casts doubt on whether observed improvements genuinely reflect corporate advancement or are merely the consequence of including lower-scoring new entrants or no development in the peer group.

Given these limitations, ESG scores cannot reliably be used to assess a company's progress in ESG performance over time. At best, they can only serve as a tool for comparison against an industry peer group (for the E and S pillar) or a country peer group (for the G pillar) at a given point in time. This limitation significantly undermines the current application of ESG scores, which is to identify leaders and laggards in the ESG environment. In the following sections, we point out these limitations and introduce a methodology that overcomes these issues, providing a more accurate representation of companies' real-time development.

2 ESG scoring methodologies

2.1 The Refinitiv's ESG score (RR)

Understanding the limitations introduced by Refinitiv's ESG scoring methodology necessitates a clear grasp of each step involved. This section delineates these steps. Refinitiv calculates its ESG score as a weighted average of the three pillar scores: Environmental (E), Social (S), and Governance (G), each determined as a weighted average of its own subpillar scores. These weights vary by sector and are transparently disclosed. The core of this process is the double application of percentile ranking which Refinitiv states to apply to avoid undue influence of outliers. To complement this section, two flowcharts are provided in Figures 2 and 3 below. More specifically, we show in Figure 3 an example of the computations of the "Resource Use" score, which is one of the three subpillars of the Environmental pillar.

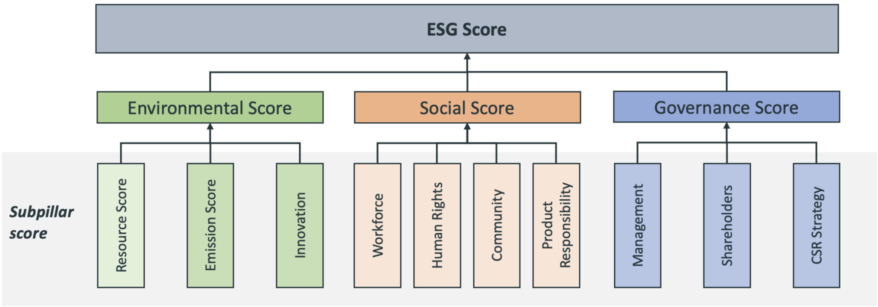


Figure 2: Flowchart of Refinitiv methodology for overall ESG score

Over 500 raw ESG-related indicators are collected for each company, categorized into industry groups. Within each industry, specific indicators - termed scoring variables - are selected and aggregated to form the final ESG score for that sector. While some indicators are common across many sectors, others are unique to a few, each influencing only one subpillar.

For each company within an industry, the raw value of its scoring variables is transformed into an indicator score using the percentile ranking methodology. These scores range between 0 and 100, with higher scores indicating superior performance compared to the other firms in the industry (for E and S pillars) or in the country (for G pillar). Missing values are assigned a score of 0. After calculating indicator scores for all variables within a subpillar, they are summed, and this aggregate is again subjected to percentile ranking to yield the final subpillar score.

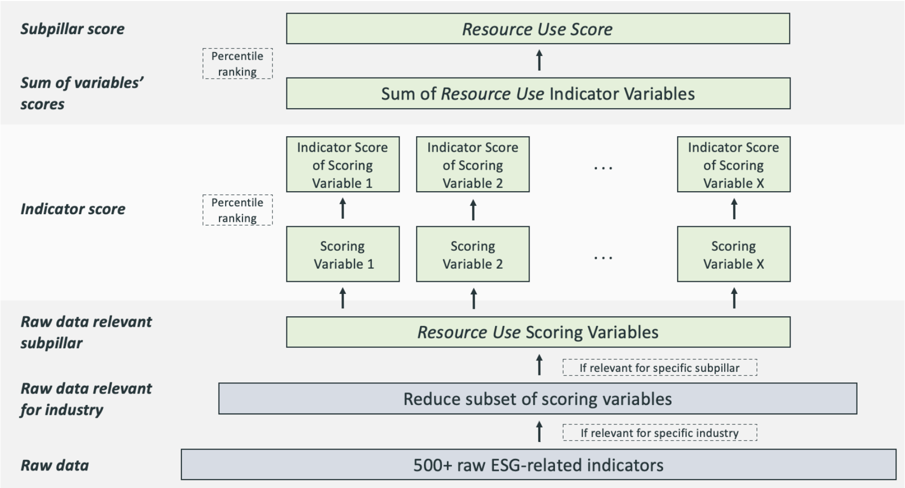

Subpillar score
Resource Use Score
Percentile ranking
Sum of variables' scores
Sum of Resource Use Indicator Variables
Indicator Score of Scoring Variable 1
Indicator Score of Scoring Variable 2
...
Indicator Score of Scoring Variable X
Indicator score
Percentile ranking
Scoring Variable 1
Scoring Variable 2
...
Scoring Variable X
Raw data relevant subpillar
Resource Use Scoring Variables
If relevant for specific subpillar
Raw data relevant for industry
Reduce subset of scoring variables
If relevant for specific industry
Raw data
500+ raw ESG-related indicators

Figure 3: Flowchart of Refinitiv methodology for the subpillar Resource Use score.

The percentile ranking methodology exhibits significant drawbacks. It effectively flattens the performance differences between companies, leading to a loss of relative performance insights. Moreover, it tends to artificially inflate the variance in indicator scores. To illustrate, consider n ranked values of a quantitative variable X, where $x_i = i/n$, $i = 1$, for $i = 1, \dots, n$. Given that $\sum_{i=1}^{n} i = n(n+1)/2$ and $\sum_{i=1}^{n} i^2 = n(n+1)(2n+1)/6$, the mean and the variance of X can be explicitly calculated as

$$\mu_X = \frac{1}{n} \sum_{i=1}^{n} \frac{i}{n} = \frac{1}{n^2} \frac{n(n+1)}{2} = \frac{n+1}{n} \frac{1}{2} \tag{1}$$

which converges to $1/2$ as $n \to \infty$. As far as the variance is concerned, note that

$$\sigma_X^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{i}{n} - \frac{n+1}{2n} \right)^2 = \frac{1}{12} \frac{n^2-1}{n^2} \tag{2}$$

which converges to the value $\frac{1}{12}$ as $n \to \infty$. Indeed, X converges in distribution to a Uniform distribution over the interval $[0, 1]$, denoted as $U(0, 1)$. It is important to note that for a distribution with density f concentrated on $[0, 1]$, the $U(0, 1)$ distribution is characterized by having maximum entropy (Dudewicz and Van Der Meulen (1981)), signifying the greatest dispersion of information. Concerning the variance, for distributions over the interval $(0, 1)$, it is maximized when the distribution is polarized at the extremes. In contrast, for unimodal distributions, whether symmetric or asymmetric, the variance is generally lower compared to the $U(0, 1)$ case. For a more detailed understanding, the Beta distribution (Johnson, Kotz, and Balakrishnan (1995)) offers a comprehensive generalization of the Uniform distribution over the unit interval, with its density being

$$f(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, \quad x \in (0,1), \quad \alpha > 0, \quad \beta > 0, \tag{3}$$

where $B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$ is the Beta function (Abramowitz, Stegun, and Romer (1988), Section 6.2). Note that the $U(0,1)$ distribution is a specific instance of Beta$(1,1)$, and its variance can be expressed by the formula $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. Figure 4 illustrates some examples along with their respective variance values. It is noteworthy that a variance exceeding $1/12$ is only achieved in cases of highly polarized distributions.

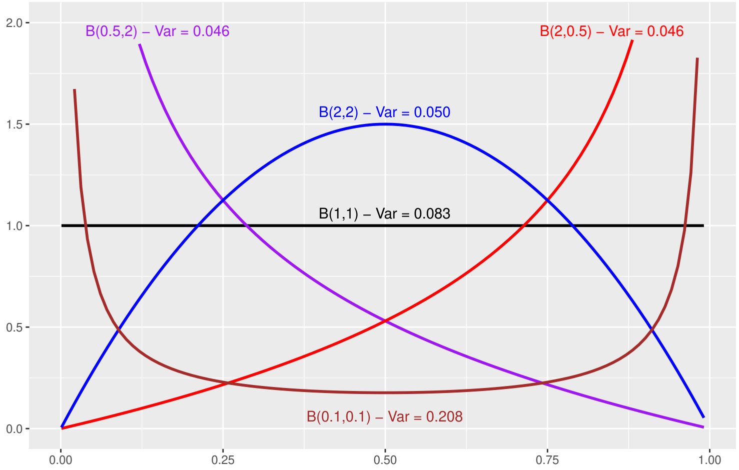


Figure 4: Beta distribution densities and their variance.

2.2 The Performance Ratio Approach (PR)

ESG score construction is a complex, multi-objective problem requiring the aggregation of diverse values. While various methods have been proposed, as reviewed in Bentley and Wakefield (1998) with an emphasis on criteria like range independence and importance, we suggest using the so-called Performance Ratio (PR) approach. This method, aligned with method 5 in Bentley and Wakefield (1998), utilizes normalized measurements instead of rankings for computing ESG scores. A similar approach has been also suggested in Roncalli (2023).

For each company in a given sector, each scoring variable raw value is mapped into an indicator score using

$$\text{PR} = \frac{Value - Min}{Max - Min} * 100,$$

where Min and Max represent the minimum and maximum values for that variable in the sample of peer companies, and $Value$ is the value of the variable for the company under analysis. This ratio-based approach eliminates range-dependence and allows for weighted aggregation of different variables.

To compute the E, S, and G pillars, for each company the performance ratios of all scoring variables within a subpillar are first calculated and then averaged. For boolean variables, this system awards a company with a score of 100 for positive performance (True for positive polarity variables and False for negative polarity ones), and 0 for negative or missing performance. Subsequently, subpillars are combined into a weighted linear combination, with weights assigned based on the number of raw variables in each subpillar, essentially creating a weighted mean. E, S and G pillars are then computed as weighted means of subpillars and similarly aggregated into the final ESG score.

Table 1 showcasts the scoring differences between Refinitiv and PR approaches at the variable level. Consider ten initial companies (C1 to C10) with raw performance values in column 2, for which Refinitiv (RR1) and PR (PR1) scores are computed. Notably, C9 and C10's underperformance in this positive-polarity variable is obscured by RR1's uniform 0-100 distribution, but properly measured in PR1's notable gap post-C8. When companies C11 to C15 join in a subsequent period, with performances between C8 and C9, and the scores are re-calculated for C1 to C15 (RR2 and PR2), RR2 scores increase for companies outperforming C11 to C15 and decrease otherwise. Contrarily, PR1 and PR2 show no change, as the PR approach's score adjustments hinge solely on the sample's Min or Max values[1].

In summary, the PR approach's scoring is less influenced by other companies in the sample. For boolean variables, the score remains constant, while for numeric variables, it is solely based on the Max and Min performers.

[1] The PR indicator would exhibit variation only if the Min or the Max value changes, not every time a company joins the universe.

	Scoring V.	RR1	RR2	RR2-RR1	PR1	PR2	PR2-PR1
C1	1.997	95	96.67	1.67	100.00	100.00	0.00
C2	1.899	85	90.00	5.00	94.55	94.55	0.00
C3	1.700	75	83.33	8.33	83.45	83.45	0.00
C4	1.621	65	76.67	11.67	79.09	79.09	0.00
C5	1.619	55	70.00	15.00	78.96	78.96	0.00
C6	1.602	45	63.33	18.33	78.03	78.03	0.00
C7	1.527	35	56.67	21.67	73.84	73.84	0.00
C8	1.403	25	50.00	25.00	66.92	66.92	0.00
C9	0.400	15	10.00	-5.00	11.13	11.13	0.00
C10	0.200	5	3.33	-1.67	0.00	0.00	0.00
C11	0.984		43.33			43.63	
C12	0.953		36.67			41.92	
C13	0.926		30.00			40.39	
C14	0.740		23.33			30.04	
C15	0.669		16.67			26.11	

Table 1: Comparing Refinitiv' (RR) and PR approaches. In black, companies considered since time 1, in red companies added at time 2. Column 1: company identifier, Column 2: Raw values of a positive-polarity scoring variable, Column 3: Refinitiv's percentile ranking for companies at time 1, Column 4: Refinitiv's percentile ranking for companies at time 2, Column 5: Refinitiv's score difference, Column 6: Performance Ratio's approach at time 1, Column 7: Performance Ratio's approach at time 2, Column 8: Performance Ratio's score difference.

3 Data

Our analysis focuses on the Refinitiv's universe of companies in three key industry groups, selected due to their sample size and relevance to climate change: (i) Machinery, Tools, and Heavy Vehicles, Trains and Ships, (ii) Oil & Gas, and (iii) Chemicals. In particular, we analyze ESG data spanning five years, from 2017 to 2021, to observe temporal changes. For each of these sectors, we have compiled five separate datasets, one for each year, including only those companies featured in Refinitiv's sector sample for the respective year. Our dataset encompasses not only the overall ESG scores for each sector but also the scores of individual pillars and subpillars. Moreover, it includes the raw variables of individual indicators and their corresponding scored values for each scoring variable, offering a detailed insight into the ESG performance dynamics. Table 2 presents the frequency of companies in the sample by sector and year, illustrating the influx of new companies into the study's universe over time.

Sector	2017	2018	2019	2020	2021
Machinery	284	335	394	480	529
Oil	191	201	219	244	248
Chemicals	170	197	228	273	306

Table 2: Number of companies in the sample of peers by sector and year.

Furthermore, companies in the original 2017 sample have higher market capitalization compared to those incorporated later as shown in Table 3.

Year	*Mean*	*Q1*	*Median*	*Q3*
2017	9.97	1.31	4.23	8.74
2018	4.37	0.62	1.36	2.12
2019	2.97	0.36	1.41	3.78
2020	2.21	0.15	0.44	1.89
2021	2.35	0.22	1.28	2.95

Table 3: Summary statistics of 2021 market capitalization of companies in the Machinery sector by year of entrance in the sample.

This size bias has already been explored in the literature by authors such as Drempetic, Klein, and Zwergel (2020) or more recently by Dobrick, Klein, and Zwergel (2023).

4 Decomposition of ESG pillar variation

Refinitiv's percentile-based methodology means that a company's score fluctuates based on its relative standing against its peers. From 2017 to 2021, the changes in the scores of one company can be decomposed into three factors: (i) the company's own ESG improvements, (ii) the variation in the ESG performance of the existing peers, (iii) the introduction of new companies into the peer group.

The following analysis focuses on dissecting the variation in these scores. Our focus is primarily on the changes in the Environmental (E) pillar scores within the Machinery sector, spanning from 2017 to 2021. Similar results apply for the E and S pillars for the three sectors and we can reasonable assume that they hold for the G pillar where the peer group composition is done at country level.

The findings are illustrated in Figure 6.[2] We begin by examining a specific case from an exemplary companies in our sample, as illustrated in Figure 5. To comprehend the details of this figure, let us first focus on the blue boxes:

The outer boxes labeled "Score 2017" and "Score 2021" display the E pillar scores for one company in the years 2017 and 2021 provided by Refinitiv, respectively. Our ability to replicate Refinitiv's methodology allows us to calculate the expected E pillar score for this company in 2021, assuming no improvement within the peer group and no addition of new companies with low E pillar scores. This expected score is represented in the second box from the left, named "Score 2021 ex new ex others" and highlights the company's actual progress from 2017 to 2021 excluding new companies and other companies in the peer group. The third box from the left, "Score 2021 ex new," represents the anticipated 2021 score for the company under the assumption that other companies did make progress (either improvement or decline) but no new company was added to the universe of the industry group under analysis.

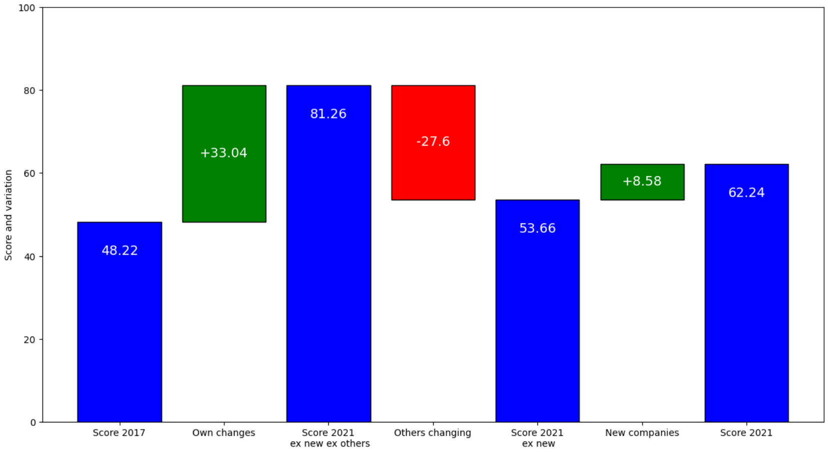


Figure 5: Decomposition of the sources of the variation in the E pillar score from 2017 to 2021 for one company in the Machinery sector, computed with the RR methodology.

[2] It's noteworthy that we observed analogous trends in the Social (S) pillar and the other two sectors, details of which are available upon request.

This analysis brings to light two critical observations. The first green box on the left, labeled "Own Change" displays the actual improvement of a company from 2017 to 2021. This reflects the intrinsic progress at the company level, not adjusted for any changes in the universe. The significant increase of +33.04 points is somewhat mitigated when compared to the company's peers, which are also making substantial improvements. This is indicated by the red decrease to -27.6 called "Others changing". Therefore, from this perspective, the company's improvement adjusts from 48.22 to 53.66, considering the development of its peers.

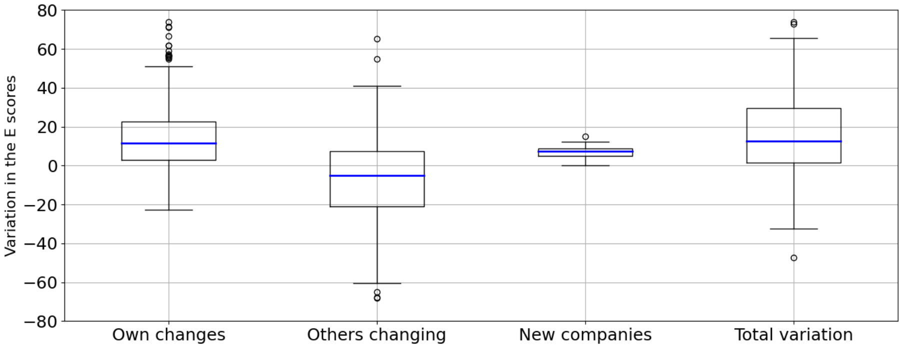
The second key insight is found in the second green box labeled "New companies". This box indicates the score increase driven by the expansion of the E peer group universe. The mere addition of new companies, most of which have poorer Environmental performance, provides an artificial boost to this company's score, pushing it up to 62.24. This scenario reveals two main mechanisms at play: 1) the company's actual development is obscured, and 2) an expansion in the E peer group universe artificially inflates the score.

Understanding these dynamics reveals a troubling truth: E pillar scores of individual companies might be misleading and unreliable as indicators of genuine progress. At best, they serve a marginal role in facilitating comparisons among peers, falling short of providing meaningful insights. These insights are also valid for the S pillar (i.e. results available upon request) and consequently also for the G pillar when considering the country level peer groups.

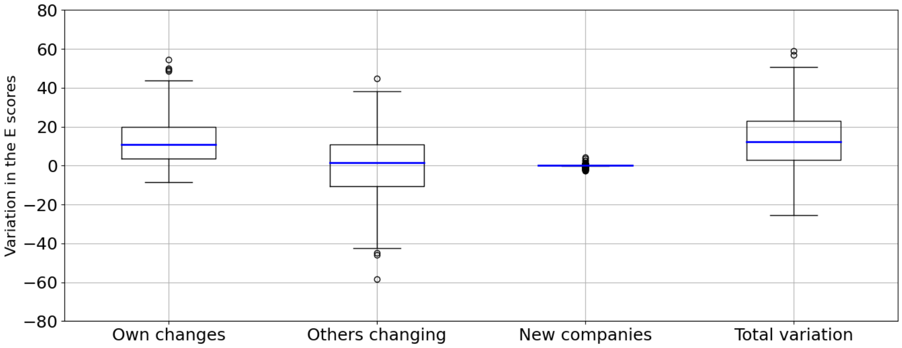
Figure 6a substantiates our theory: Figure 6 illustrates the changes previously discussed for an individual company using the RR methodology, now extended to encompass all companies in a particular sector. Figure 6a presents these changes, while Figure 6b offers a similar comparison but utilizes the PR methodology. The first three box plots in Figure 6a correspond to the two green and one red box previously examined variations in the individual case. The final box plot, labeled "Total Variation", depicts the variations in the peer universe observable to investors when analyzing the Refinitiv scores – for instance, in the example above for a single company, this would be a change from 48.22 to 62.24. Here, we consider the universe of companies in the sector under investigation in the initial year.

Focusing specifically on the RR methodology in Figure 6a, it is noteworthy that the "Own Change" box (representing intrinsic company-level changes from 2017 to 2021) and the "Total Variation" box (indicating the change observed using the final Refinitiv scores from 2017 versus 2021) are not similar in size, with the median being higher for "Total Variation". This supports our earlier observation that the evolution of a single company is not accurately reflected in the RR scores. Furthermore, while the median of "Others Changing" is negative, suggesting a decrease in scores due to alterations in the peer group, it is particularly important to note that the whole distribution of "New Companies" is positive (or zero), resulting in an artificial inflation of scores of all existing firms as the universe of companies expands.

In contrast, the lower box plots in Figure 6b, which are based on the PR methodology outlined in Section 2.2, display a different trend, with no apparent



(a) Decomposition of the sources of the variation in the E pillar score from 2017 to 2021 for the Machinery sector sample in 2017, computed with the Refinitiv's methodology.



(b) Decomposition of the sources of the variation in the E pillar score from 2017 to 2021 for the Machinery sector sample in 2017, computed with the PR methodology.

Figure 6: The four boxes show the distribution of (i) the total variation in the period for the E pillar attributable to changes of the single companies, (ii) the total variation in the period for the E pillar attributable to changes of the peers, (iii) the total variation in the period for the E pillar attributable to the enlargement of the universe of companies, (iv) the total variation in the period for the E pillar. The upper boxplots report the differences in scores computed using Refinitiv's methodology, while the bottom boxplots report the differences in scores computed using the PR methodology.

obscuration of scores. Here, the "Own Change" and "Total Variation" box plots are strikingly similar in both size and median value. Additionally, while the "Others Changing" category results in minor positive and negative variations in the score, its median hovers around zero, mirroring the effect seen in the "New Companies" category. This alignment around the zero mark indicates that the introduction of new companies into the analysis has a negligible impact on the overall scores. Furthermore, changes in the scores of other companies exert a limited effect, underscoring that the primary driver of score variation is the companies' own E pillar disclosures and advancements.

The primary goal of this analysis is not to establish the PR method as a superior approach for E pillar scoring. Rather, its focus is to highlight that the double percentile ranking method, detailed in Section 2 and depicted in Figure 3, results in an artificial obscuration of scores for companies that have been in the sample for a longer duration. This phenomenon is noteworthy because it can be effectively addressed and mitigated by employing the PR method. This insight underscores the potential for more accurate and representative E pillar scoring methodologies.

Investors relying on these scores might be led to believe that companies are genuinely improving their E pillar performance by making positive changes. However, this perceived improvement could merely be a result of the inclusion of lower-performing companies in the E pillar database, rather than actual advancements made by the longer-standing companies in the sample.

To provide a more detailed and quantitative explanation, we examine the impact of new companies entering the sample on E pillar grades (A, B, C, D), which are assigned based on score values[3]. This analysis differentiates between the two methodologies. Figure 7 illustrates the variance in classifications using the RR method (left) versus the PR method (right).

For each methodology, we calculate two confusion matrices[4]. The first matrix tracks the grade movements during the 2017-2021 period. The second matrix is similar, but the 2021 grades are recalculated, excluding new companies that entered the sample between 2017 and 2021. Consequently, Figure 7 displays the discrepancies between these two matrices (Computed as Matrix 1-Matrix 2), highlighting how the inclusion or exclusion of new entrants affects the grading shifts under each methodology.

A positive number indicates that, when the entry of new companies into the database is accounted, more companies received the grade indicated in the column than when the entry of new company is excluded from the computations. A negative number is interpreted in the opposite way. The interpretation is the following: consider the first row of the left matrix, which is based on the RR methodology. 7 companies that received an 'A' grade in 2017 were able to retain their 'A' grade in 2021 when we account for new entries of companies. However, if the sample had been made up of the companies that were in the sample in 2017, then 5 of these seven companies would have obtained a 'B' grade and 2 of them would have received a 'C' grade. Similarly, the second row suggests that 16 companies that had received a 'B' grade in 2017 were upgraded to 'A', but without the entrance of new companies, they would have either confirmed their 'B' (in 10 cases), or even worsened to 'C' (4 cases) and 'D' (2 cases). While the left matrix shows a significant inflationary effect due to the entrance

[3] Grade A is assigned for scores larger than 75, grade B is assigned for scores between 50 and 75, grade C is assigned for scores between 25 and 50, grade D is assigned for scores lower or equal than 25. Refiniv defines a more granular scale with +/-, which we consider within the letter. Therefore, for simplicity we consider A the grades from A+ to A-.

[4] These matrices are not included in the paper but are available upon request

of new companies, the right matrix does not exhibit this tendency. The only minor difference in the PR confusion matrix are due to the variation of the Max or Min due to companies joining the sample during the time period under consideration.

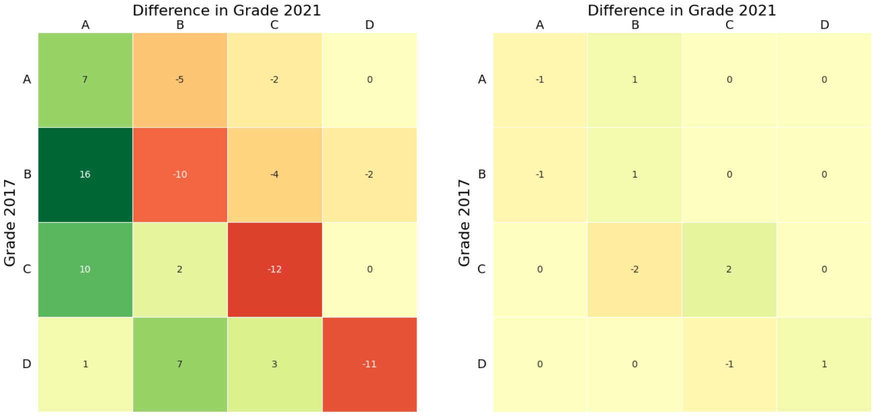


Figure 7: Difference in class switches in the 2017-2021 period due to new companies entering in the sample according to the Refinitiv's methodology (left matrix) and Performance Ratio (right matrix) for the E pillar.

4.1 New companies and missingness

Refinitiv's expansion of its ESG company universe impacts individual scores, influenced by new sector-classified companies. The introduction of companies follows a systematic approach based on size and location, potentially introducing inflation. Refinitiv's report (page 6, Refinitiv (2023)) reveals its initial coverage included major indices like SMI, DAX, CAC 40, with subsequent additions like DJ, STOXX, and MSCI World (in 2008). Regular additions are made yearly, contingent on gathering *sufficient*[5] public information for a comprehensive ESG assessment. Companies, unable to opt out of scoring, may provide additional information to enhance their scores. Initially, many scoring variables may be missing, but there's an incentive for companies to disclose positive information over time to improve their comparative scores.

Refinitiv's selection criteria for expanding its ESG universe lead to predictable outcomes, evident in Table 3. Specifically, companies in the original 2017 sample have higher market capitalizations in 2021 compared to those included later. Larger firms, with more resources, are generally more adept at implementing ESG projects and disclosing more information. This link between company size and ESG scores is well-documented in research, such as Drempetic

[5] According to Refinitiv customer service sufficient information refers to the level of detail and coverage required to provide a comprehensive assessment of a company's ESG performance.

et al. (2020). Figure 8 illustrates a higher incidence of missing scoring variables in firms added to the sample post-2017 compared to those included in 2017. Moreover, Figure 9 demonstrates a clear trend: companies earlier in the sample are more likely to attain higher grades (A or B) in the E pillar. This is contrasted by higher D-ratings (marked in red) among companies added in later years.

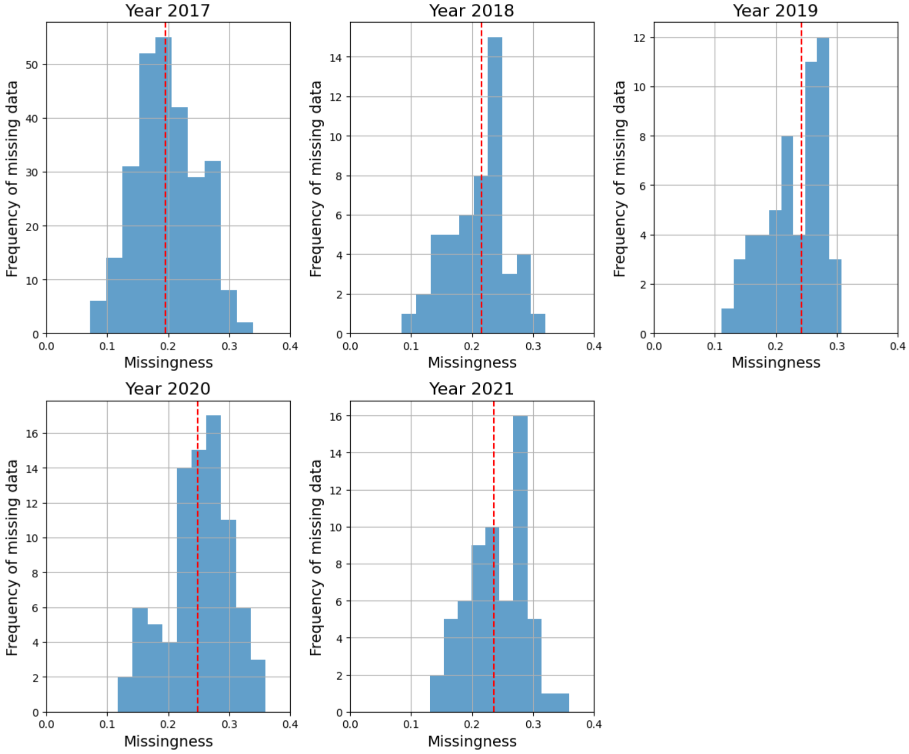


Figure 8: Histograms of the percentage of missing scoring indicators for the overall ESG score (see Figure 3) in 2021 Refinitiv data for the Machinery sector, distinguishing by year of entrance in the database. Median missingness is shown by the red dashed lines. The plots refer to the Machinery sector. Similar results apply for other sectors and are available upon request.

Our analysis indicates that new, typically smaller entrants with lower ESG performance tend to fare poorly relative to established companies that have had more time to disclose information. This suggests that expanding the peer sample may artificially inflate the ESG scores of existing companies, as they are compared with these newer, lower-performing entrants, which often have a higher degree of missing data.

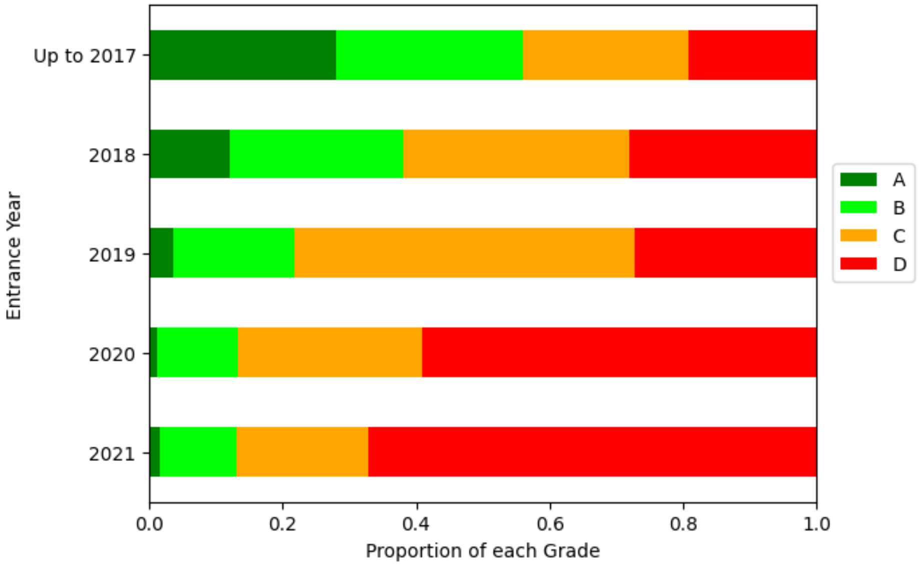


Figure 9: Distribution of the E pillar grade in year 2021 for companies in the Machinery sector, dividing by year of entrance. We find similar results for the S pillar score which are available upon request.

4.2 Change in existing companies' disclosures

The expansion of the peer group is likely to initially elevate the scores of companies already in the database. However, as competition intensifies within this group, a natural compression of scores may occur. Consider a scenario where a company was a top performer in the initial year. If, in subsequent years, it does not enhance its disclosed values while its peers do, this company will struggle to maintain its former relative standing and consequently, its score will diminish compared to the first year. In a context where peer companies are progressively improving their disclosures, a firm must not only match but exceed these improvements to enhance or even just preserve its relative position and score.

5 Comparing Methodologies: Refinitiv versus PR

In the following section, we present a detailed comparison of our PR approach with the Refinitiv methodology, concentrating on individual pillars. Our analysis primarily targets the Environmental (E) and Social (S) pillars, as these allow for a more feasible comparison due to the consistent sector-based company universe in both methodologies. Additionally, we have included a comparison of the Governance (G) pillar in the appendix. The disparity in G pillar scores is notably more distinct between Refinitiv and PR methodologies, largely due to a transition from percentile ranking to PR and a variation in the company target universe, as detailed in Appendix A.

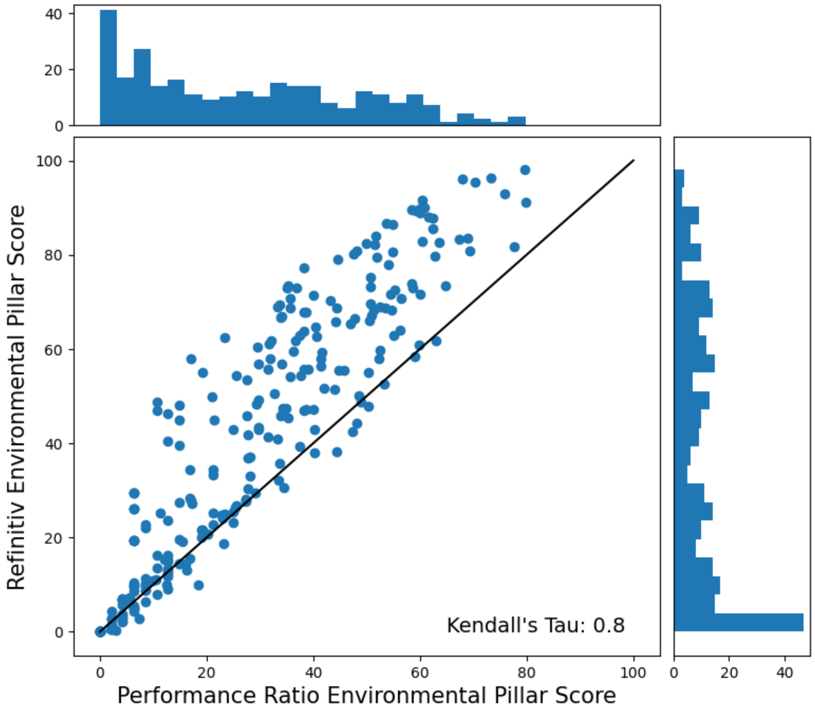


Figure 10: Scatterplot of the E pillar score computed with the Performance Ratio methodology versus Refinitiv's E pillar score. The plot refers to the Machinery sector in the year 2017.

The scatter plots in Figures 10 and 11 compare our PR methodology (x-axis) with Refinitiv's methodology (y-axis) for E and S pillar scores in the Machinery sector for 2017. Similar trends are observed in other sectors and years, details of which are available upon request. The plots also include histograms showing the marginal distribution of both variables.

We observe a strong correlation between the two methodologies, with Kendall's tau values of 0.8 and 0.88 for the E and S pillars, respectively, in the Machinery sector. However, there is a discernible trend: companies scoring high with Refinitiv's methodology also do well in ours, but generally with lower scores. This is attributed to Refinitiv's more dispersed scoring across the 0-100 range, owing to the double application of percentile ranking. In contrast, our PR methodology sets a higher bar for top scores, requiring a company to outperform in all scoring variables to achieve a perfect score of 100. The lack of companies obtaining top scores with the PR methodology indicates that ESG leaders, although

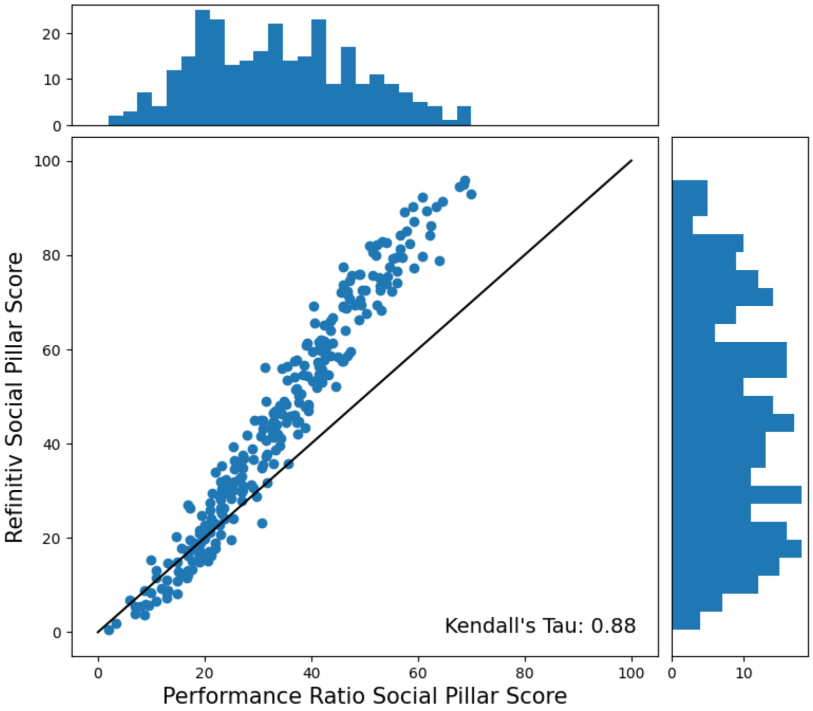


Figure 11: Scatterplot of the Social pillar score computed with the Performance Ratio methodology versus Refinitiv's Social pillar score. The plot refers to the Machinery sector in the year 2017.

performing better than their peers both in absolute and in relative terms, still have significant room for improvement.

Diving deeper, Figure 12 shows histograms comparing E and S pillar scores calculated using Refinitiv's and PR methodologies. For 2017 and 2021 (with similar trends in other years and sectors), PR's method often leads to notable score reductions, occasionally over 30 points, while increases are generally minor, rarely above 10 points. This indicates that Refinitiv's percentile ranking method tends to inflate E and S pillar scores, a limitation that becomes evident when the aggregation method is refined and corrected, resulting in significantly lower E and S pillar scores.

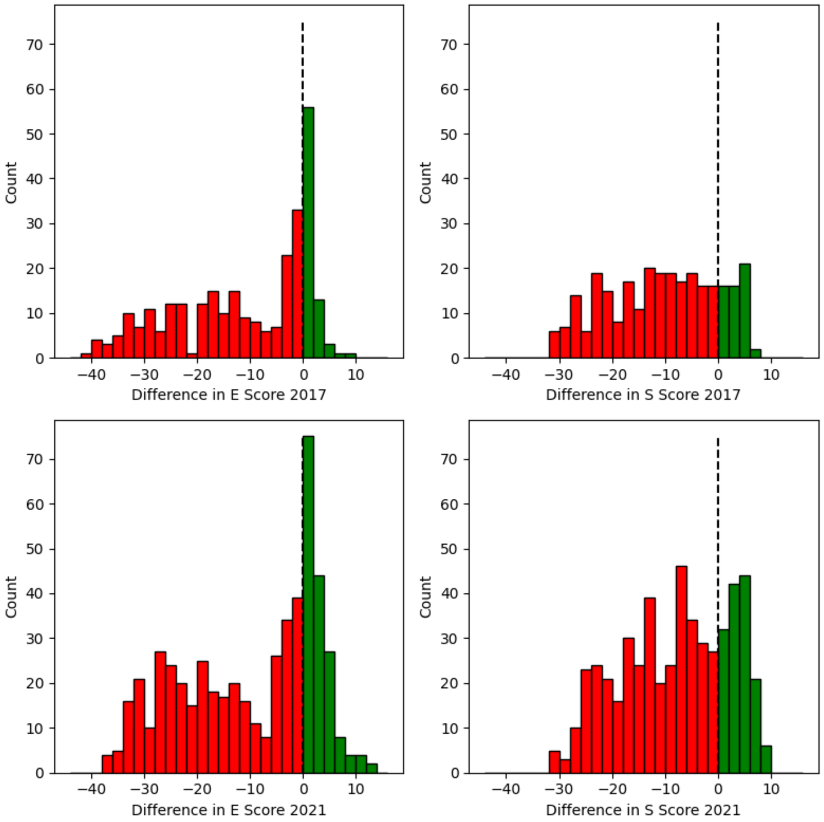


Figure 12: The plots display the histograms of the difference between the E pillar scores and the S pillar scores computed with the PR methodology and those computed with the Refinitiv's methodology for the Machinery sector.

6 Conclusion

In this paper, we scrutinize the efficacy and accuracy of Refinitiv's percentile ranking in ESG scoring, probing whether apparent improvements in scores truly reflect corporate advancement or are influenced by the entry of lower-scoring new companies and the relative performance with respect to the peer group universe. Our analysis uncovers a positive inflation in Refinitiv's approach, where the addition of companies with limited information distorts ESG performance portrayal. In response, we propose a 'performance ratio' based scoring system, offering a more precise measure of actual performance, especially in evaluating ESG score leaders.

Our research focuses on three critical sectors: Machinery, Tools, Heavy Vehicles, Trains and Ships, Oil & Gas, and Chemicals. These sectors were selected for their large sample sizes, their relevance to climate change issues, and their active engagement in E and S pillar measures. Spanning five years, from 2017 to 2021, we conduct a thorough examination of ESG data, encompassing overall scores, scores across individual pillars and subpillars, and both the raw and scored values of each indicator. This extensive dataset provides a detailed perspective on the evolution of E and S pillar performance across these sectors, which should also reasonably hold for all sectors.

Our analysis reveals a notable correlation between Refinitiv's methodology and our proposed PR approach. A key finding is that while companies scoring high in Refinitiv's system also rank well in our system, without being positively inflated by the entrance of new companies, and negatively deflated by the 'inner competition'. Our deep dive into score distributions consistently shows that Refinitiv's method tends to produce inflated scores, especially for top performers.

In conclusion, our research underscores the imperative for a more robust ESG scoring system. The 'performance ratio' methodology we propose not only explicitly deals with the inflation present in existing systems but also seems rather insensitive to outliers which was stated as the reason from Refinitiv to introduce percentile ranking methodologies. As the importance of ESG continues to escalate in the corporate world, it is essential that scoring methodologies advance in tandem to truly and accurately reflect the sustainable advancements of corporations, reflecting the information content of the variables aggregated to build such scores. Further research could focus on alternative methods, such as the ones suggested also by Roncalli (2023), as well as improving the aggregation and information content of scores built from both boolean and real variables.

A Comparing Methodologies: Refinitiv versus PR for the G pillar

The contrast in G pillar scores between Refinitiv and PR methodologies is more pronounced due to a shift from percentile ranking to PR and a different in the company target universe, as illustrated in Figure 13. Refinitiv's country-level comparison differs from PR's sectoral assessment, aligning with Environmental and Social pillars, leading to pronounced score disparities. This results in a lower yet significant Kendall's tau correlation (0.31). For consistency, we focus on the E and S pillars, where the sector defines the company universe annually, allowing quantification of the impact of new sector entrants. Instead for G pillar the reference universe is at country level and less accessible. However, our work shows the limitations of percentile ranking in Refinitiv's approach which should also hold for the G pillar.

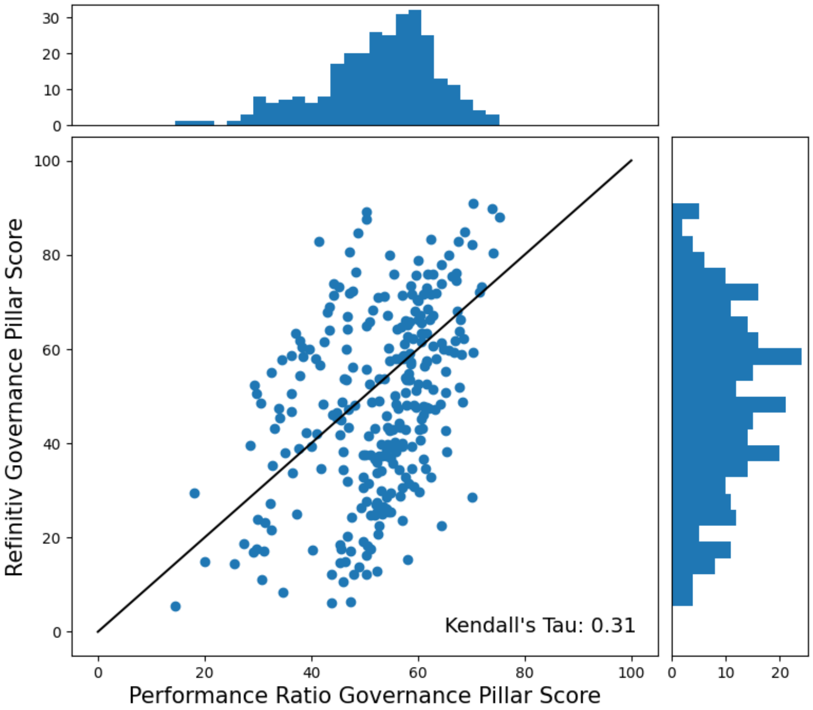


Figure 13: Scatterplots of the Governance pillar score reproduced with the Performance Ratio methodology versus Refinitiv's Score. The plots refer to the Machinery sector in the year 2017.

References

Abramowitz, M., Stegun, I. A., & Romer, R. H. (1988). *Handbook of mathematical functions with formulas, graphs, and mathematical tables.* American Association of Physics Teachers.

Bentley, P. J., & Wakefield, J. P. (1998). Finding acceptable solutions in the pareto-optimal range using multiobjective genetic algorithms. In *Soft computing in engineering design and manufacturing* (pp. 231–240).

Berg, F., Fabisik, K., & Sautner, Z. (2020). Rewriting history ii: The (un) predictable past of esg ratings. *European Corporate Governance Institute–Finance Working Paper*, *708*(2020), 10–2139.

Dobrick, J., Klein, C., & Zwergel, B. (2023). Size bias in refinitiv esg data. *Finance Research Letters*, *55*, 104014.

Drempetic, S., Klein, C., & Zwergel, B. (2020). The influence of firm size on the esg score: Corporate sustainability ratings under review. *Journal of business ethics*, *167*, 333–360.

Dudewicz, E. J., & Van Der Meulen, E. C. (1981). Entropy-based tests of uniformity. *Journal of the American Statistical Association*, *76*(376), 967–974.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous univariate distributions, volume 2* (Vol. 289). John wiley & sons.

Refinitiv. (2023). *Environmental, social andgovernance scores from LSEG* (No. May). Retrieved from `https://www.lseg.com/content/dam/data-analytics/en`$_u$*s/documents/methodology/lseg − esg − scores − methodology.pdf*

Roncalli, T. (2023). *Lectures on hedge fund strategies.* Retrieved from `http://www.thierry-roncalli.com/download/HSF-Lectures.pdf` (Accessed: [04.12.2023)

Sahin, Ö., Bax, K., Czado, C., & Paterlini, S. (2022). Environmental, social, governance scores and the missing pillar—why does missing information matter? *Corporate Social Responsibility and Environmental Management*, *29*(5), 1782–1798.