ESG ratings: why can’t the raters agree?

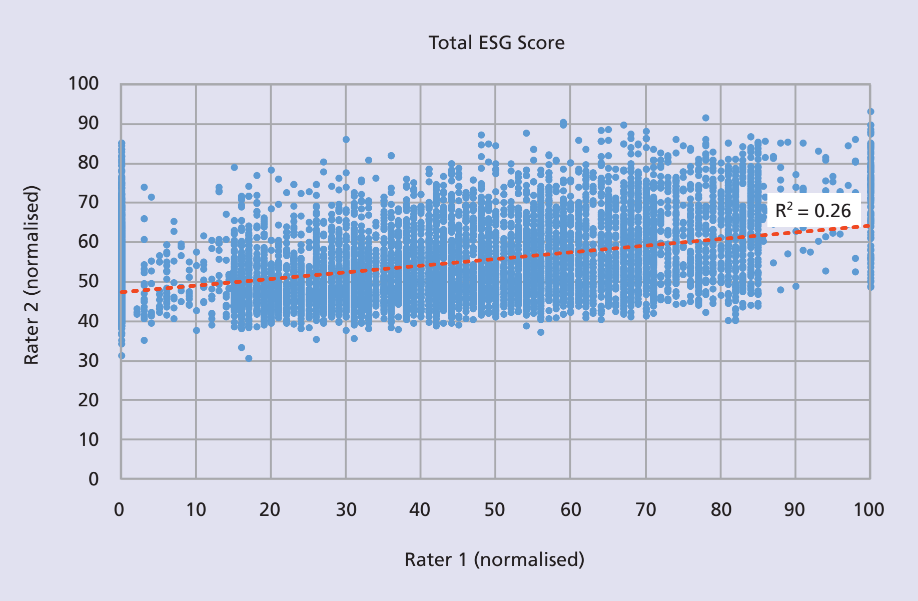
Lara Kesterton
ESG Analyst
Vontobel Asset Management

Following years of expectation, sustainable finance has become mainstream and is growing significantly. According to one measurement, at the end of 2018 there were already some 18 trillion US dollars invested in line with ESG integration approaches, an increase of 69% versus the end of 2016.[1]

With this tailwind, rating agencies that assess ESG factors to help investors make informed decisions on sustainable investing are booming, with more than 125 different agencies established worldwide[2]. These raters assess a number of different metrics, applying their own proprietary magic formula for how to aggregate, weigh, and come up with an overall number or grade. Akin to a credit rating score, this might give the impression of a consensus-drawn evaluation derived from hard facts and defensible figures, but these grades mask layers of subjectivity and hidden biases. In fact, approaches, and therefore results, of ESG raters differ widely as Figure 1 illustrates.

Recent academic research performed similar analysis more broadly, finding a correlation coefficient of around 0.49[3] when comparing the scores of different leading ESG raters. To put this into context, this is contrasting to a coefficient of 0.96[4] (indicating strong agreement) for credit rating agencies, where the industry landscape and approaches are much more consolidated and long-standing. The research confirms that ESG rating agencies agree neither on what constitutes good ESG practice, or on who is good or bad at it. Particularly, there was a stark disagreement in the tails of the ratings, which is notable as many investors use these results to create best-in-class portfolios or avoid worst-in-class performers.

Figure 1: Comparison of ESG scores between two leading ESG rating agencies



Source: Vontobel Asset Management as of November 15, 2019. Company universe based on rater 1 universe

“Rating agencies that assess ESG factors to help investors make informed decisions on sustainable investing are booming.”

One underlying problem is that ESG raters serve various responsible investing interests. In practice, the raters usually go about the rating process by developing proprietary methodologies to rank and score companies on the panoply of ESG issues.

ESG raters take data from multiple different sources and languages and use models to clean, organise, and weight these diverse data points to create comparability and to flag risks. The scoring models used by ESG raters of course have their merits by giving structure to decision-making, but they are also at risk of giving the impression of scientific rigor, when in fact ESG practice is still an art and subjective, rather than being a science.

The 10 challenges of ESG ratings

1. Material factors

These affect the decision as to which ESG topics should be included in the model. While greenhouse gas emissions will be commonly assessed, indigenous rights, employee organisations, or lobbying might be considered more niche topics for assessment and only scored by a few. The number of data points evaluated by raters vary from 10 to >400, although there is good evidence that counting too much merely weakens the real target signal.[5]

2. Measurement

Raters use different metrics to evaluate a topic. For example, to evaluate employee health and safety, raters choose from 20 different data points to score this topic.[6] Some academic research found this to be the dominant reason for rater divergence.[7] Peeling back the layers of what gets measured, the raw underlying data is more inconsistent than you might think.

3. Data quality

How defensible is the ESG data? Is it pure marketing information, as non-financial information is not required to be certifiable or defensible in the same way that financial statements are? Frequently, metrics supplied by companies are patchy, inherently backward looking, and tend to fall into “good news” storytelling. Some raters exclude data provided by the company itself – which can naturally be a rich data source. Similarly, as ESG metrics are frequently qualitative, raters must choose how they interpret and score descriptive matters.

4. Gaps treatment

It is common for companies not to report on all indicators. Different statistical tools can be used to fill the gaps with widely different outcomes.[8] Interestingly, a few studies found larger firms experience more disagreement in their scores, suggesting again that more data points can lead to more disagreement between raters. An active investor with good relations with the firm can sometimes overcome data gaps by direct dialogue.

5. Timing aspects

The frequency with which raters evaluate a company can have a material bearing on discrepancies between scores. An annual review is not uncommon, but also time gaps of two years between the latest updates of different raters may exist.

6. Rater bias

The rating houses have a natural slant, e.g., a focus on best-in-class, risk, momentum, and climate. It has been observed that raters based in civil-law countries are more focused on social issues, whereas common-law countries have a shareholder-centric approach and therefore have a higher focus on governance issues[9].

In addition to explicit biases, research has shown an unexplained or unconscious “rater effect”, in that when a rater is generally positive (or negative) on a company this is reflected across the board, including on unconnected indicators. This could account for 14–18% of rater disagreement.[10]

7. Weighting methodology

Next, raters need to assign how much importance to give an indicator in their model. This is largely subjective and not always transparent. Most models have indicators with little or no statistical significance – meaning they are being scored without having any real impact on the overall ESG score or any link to financial performance.[11]

8. Controversy handling

Controversy handling is the walk of the sustainability talk, and for many raters they have a high prominence in scoring. To be comparable, controversial incidents have to be evaluated for impact on society and for the business.

9. Benchmarking

As the rater translates the scoring into a final rating, an important input is also the perspective taken.

- Relative scoring is used to benchmark performance against peers. But this raises the question – what is the right peer group? Universal comparisons or against the industry peers? If the latter, again, raters choose from different industry classification systems, such as GICS, BICS, IVA industries, or perhaps an in-house division of industries. Then throw in to the mix how to treat diversified companies, and no wonder a leader in one classification can be only average in another rater’s eyes. Additionally, relative scoring can of course miss the point on sustainability if the entire industry is not addressing the issue well enough.

- Absolute scoring is the alternative approach and scores on preset ranges or optimal levels. Subjectivity creeps in on who sets the benchmark and then this leads to natural tilts away from certain industries or countries which commonly underperform in certain areas, e.g., diversity in the financial sector or on Chinese boards.

10. Aggregation of ratings

Portfolios are also scored on their average ESG rating. In truth, the average fund scores tend to be tightly clustered in a narrow spread, therefore, a top-rated fund may not have an average score notably ahead of a weak fund. At this fund level the aggregated score is even further removed from the underlying raw data and we are now in black-box territory in terms of what the scores really ought to tell you – how exposed you are to risks and whether those risks have been adequately priced in.

How to sail around these challenges

A deafening demand across the ESG industry is for companies to supply more, better quality data that is comparable. This should address a major reason for disagreement amongst raters. There are various voluntary industry and legal initiatives[12] working to create a common set of metrics on which all companies should report.

Another way to mitigate the problem is the new wave of artificial-intelligence-driven ESG ratings that are being designed to overcome human unconscious biases and normalise for size and industry skews. Other major trends are the increasing use of unconventional data sources[13] to get more impartial risk insights as well as consolidation within the rating industry. The major raters have been on a land grab in the last few years buying up smaller, niche players, suggesting a consolidation on ESG theorisation may emerge. However, at the same time, sell-side analysts have entered the space adding alternative views.[14]

An active, high-conviction manager should look beyond aggregated ratings

For the thoughtful investor, this disillusion with ratings requires looking beyond frameworks and adopting a multi-layered approach. To start with, use informative data from the ESG raters to feed an in-depth assessment to enrich fundamental equity analysis. A step-by-step process of investigation leads to a much more detailed and holistic understanding of a company, its flaws and beauty spots, but always focusing on the key issues that are really material to that company. This detailed appreciation of the top ESG risks that can impact performance is much more informative to an active investor than the specific score crunched out at the end of the rater's model. The real goal is to use ESG information to understand if the company in question has the ability to withstand its top risks in a one-to-five-year period.

Still, at some point, an investor will wish to aggregate findings to a portfolio level and this is when care must be taken to avoid losing or overlooking details when zooming out. One way to go about it is to visualise the findings on a stock level in a tile chart, which is an aggregation of the more detailed company-by-company ESG risk assessment. This way, risk concentrations are easy to spot, without losing the important details on where exactly those risks come from.

For us as an active, high-conviction equity manager, this means we conduct our own ESG analysis. We prefer an absolute perspective, setting a minimum standard to make a company investable. We put a lot of focus on controversies, which might result in a company becoming non-investable even if it passes on the average of scores. Ultimately, we concentrate on the most important risk areas to achieve a more holistic conviction on how exposed a company is to ESG factors and how well prepared it is to navigate these challenges.

[1] Voorhes, 2018.

[2] Voorhes, 2018.

[3] This is the average of the mean correlation of the following four papers. Bender, et al., 2018 found correlation between four leading raters ranged from 0.47 to 0.76 with an average of 0.59. Gibson, et al., 2019 found average correlation between six prominent raters was 0.46. Berg, et al., 2019 found a correlation range of 0.42 to 0.73 with an average of 0.61 in their assessment of five leading ESG raters. Chatterji, et al., 2016 had the lowest mean correlation of 0.3 for six well-known raters (with a range from –.012 [indicating severe disagreement] to 0.67, and only a quarter of the correlations were higher than 0.5).

[4] Berg, et al., 2019.

[5] The Sustainability Accounting Standards Board (SASB) is leading the charge on addressing this with its endeavor to create consensus on material ESG issues for each industry and sub-sector.

[6] Kotsantonis & Serafeim, 2019.

[7] Berg, et al., 2019, Chatterji, et al., 2016.

[8] E.g. do you assign the industry average (or universal or home market peer group average) or score with lowest score or use some other statistical model or not score at all? Kotsantonis & Serafeim, 2019 examines this in detail.

[9] Gibson, et al., 2019.

[10] Berg, et al., 2019.

[11] Berg, et al., 2019.

[12] EU Non-Financial Reporting Directive has required ~6,000 EU companies to publish ESG data since 2017 annual results. Plenty of other regulatory requirements come from stock exchanges (UNSSE, ESMA); international and domestic law (e.g. legislation in discussion under EU Action Plan, French Article 173, China mandatory ESG disclosure by 2020); principles frameworks (i.e. ICMM, TCFD, SDGs, GRI, UN Global Compact); or voluntary disclosure frameworks (SASB, GRI, CDSB). The alphabet soup is discussed further in Temple-West, 2019.

[13] E.g., geographic information systems data (e.g., for real estate at risk), loyalty scores and customer reviews, independent product recall data, supply chain mapping, non-government organisation reports, employee review sites and many more.

[14] Naumann, 2019.