Florian Berg, Jason Jay, Julian Kölbel and *Roberto Rigobon**

The Signal in the Noise

KEY MESSAGES

- **The information that ESG raters produce is valuable**
- **Assessing ESG performance is conceptually challenging because we need to measure contextuality, additionality, and preferences**
- **ESG raters, specialized ESG data providers, and aggregators can harness economies of scale**
- **Regulators should enforce transparency of measurement and aggregation practices to increase competition between ESG raters to incentivize improvement**

INTRODUCTION

ESG rating agencies have been under severe bombardment lately. Criticism concentrates on several fronts: what should be measured; how should it be measured; the unfortunate opaqueness of the procedures; and the severe discrepancies across different ratings for any given firm. Some voices have called for a full-blown overhaul. In fact, it is rare to find a week without someone writing a criticism of ESG ratings and ESG rating agencies in particular. Critique from politicians, all the way to John Oliver, reflects that people are dissatisfied with the current situation. Indeed, it could be said that parts of the right and left of the political spectrum have found an intersection, albeit for different reasons, in their disapproval of ESG rating agencies. Notably, The Economist (2022) has recently argued in its cover story that financial institutions should retreat from ESG to simply focus on the environmental dimension or even more specifically, just on carbon emissions – stating as a reason that there is too much noise in the signal.

We argue that abandoning ESG would throw the baby out with the bathwater. Firms' ethical behavior is essential to the health of economies, societies, and the natural environment. ESG, however flawed, is the current best effort to measure the ethical behavior of firms. Deployed in a more transparent manner, ESG data can empower investors and other stakeholders to hold firms accountable. Our research suggests that ESG data can be an important source of information for investors, and this will be even more true as we elevate the signal in the noise.

First, we document the problem, i.e., the disagreement. Indeed, the scores from different ESG rating agencies exhibit low correlations. Figure 1 presents the score of firms for Sustainalytics in the horizontal axis – the scores have been rescaled to make them comparable (i.e., normalized to have mean zero and variance 1); and the vertical axis represents the rescaled scores of the same firm in the same year given by other rating agencies (S&P, Moody's, MSCI, Refinitiv, and KLD). If the measures were highly correlated, the cloud should look like an ellipse aligned along the 45-degree line. This is clearly not the case here.

What to do with this degree of disagreement? Some argue that ESG ratings should be standardized, whereas others even go so far to say that the ratings should simply be disregarded. According to our research, both would be a mistake. Indeed, we find that ESG ratings do contain a signal. Furthermore, given the complexity of what ESG measurement entails, we believe that the only solution to gathering, analyzing, and aggregating the data runs through commercial ESG rating agencies and ESG data providers. We also do not believe that the standardization of ESG ratings would be an appropriate solution, as this would set in stone an imperfect measure, prone to be manipulated by firms and disincentivizing all research for further improvements. However, these future improvements are what ESG ratings clearly need.

IS THERE SIGNAL IN THE NOISE?

Given the disagreement, is there any signal at all in the ESG rating agency scores? The short answer is yes! Especially for the relationship between stock returns and ESG scores.

In our recent research (Berg, Kölbel and Rigobon 2022), we think of the score of a particular ESG rat-

* We offer special thanks to the Aggregate Confusion Project members (MassPRIM, MFS, AQR, Asset Management One, and Qontigo), who have financially and intellectually supported our research. The views expressed in this article are our own.

Florian Berg

is currently a research scientist at the MIT Sloan School of Management, where he cofounded the Aggregate Confusion Project.

Jason Jay

is a Senior Lecturer and Director of the MIT Sloan Sustainability Initiative, and a co-founder of the Aggregate Confusion Project.

Julian Kölbel

is Assistant Professor of sustainable finance at the University of St. Gallen, School of Finance and Center for Financial Services Innovation and a co-founder of the Aggregate Confusion Project.

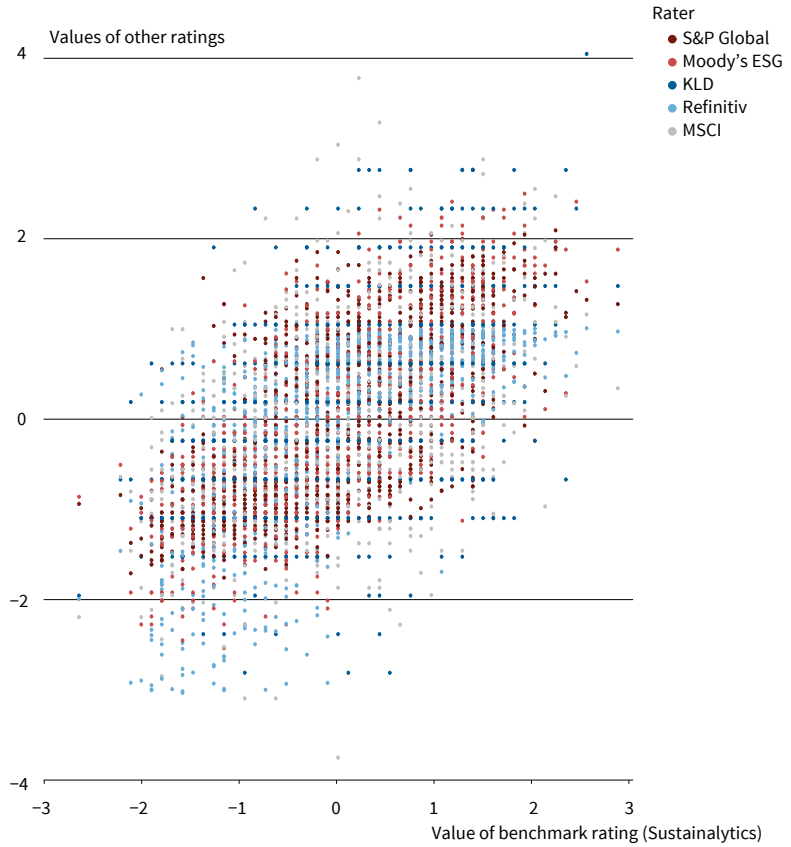
Roberto Rigobon

is the Society of Sloan Fellows Professor of Applied Economics at the Sloan School of Management, MIT, a research associate of the National Bureau of Economic Research, a visiting professor at IESA, and the founder and director of the Aggregate Confusion Project.

ing agency as the combination of some noise and an underlying true ESG performance. In our paper, we correct for the noise and find that the relationship between ESG scores and stock returns is positive and highly significant economically as well as statistically. Furthermore, we show that the reason why sometimes this relationship is hard to detect in the data, as has been the case in the literature, is precisely because the data is noisy. Think of a real-life situation, such as when you try to listen to a lecture with a lot of background noise due to construction work. The noise will drown out the signal and make the lecture harder to understand; however, the knowledge is still being imparted.

Figure 1

Correlation Between Benchmark Rating and Other Ratings



Source: Berg, Kölbel and Rigobon (2022).

To disentangle the signal from the noise, we use an instrumental variable approach where we instrument the score of one rating agency with the scores of up to seven other rating agencies. This approach consists of two stages. First, we regress one ESG rating on the other ESG ratings. Here, we do indeed find that the rating agencies are measuring something that is common across them. Second, we regress the stock return on the predicted value from the first stage while controlling for a host of financial variables, industry, and time effects. By doing so, the coefficients more than double and become statistically significant. Our results suggest that the noise implied in the ESG measures is substantial with more than 60 percent of the total score. This also means that there clearly is a signal in the ESG ratings.

SHOULD WE MEASURE CO_2 ONLY?

Should we concentrate exclusively on CO_2 measures and disregard the rest? The short answer is no.

This recommendation implicitly assumes that the environmental dimension is better measured than the social or governance dimension. There are many reasons that suggest this assertion is incorrect. We discuss those below. However, even if it were the case that CO_2 emissions are measured better than social aspects, such as discrimination of historically disadvantaged groups, the recommendation of measuring the first but not the second still does not make sense. Often the most relevant issues are hard to measure. Arguing that because something is difficult to measure it should be disregarded is questionable at best.

It is unwise to limit ourselves to only the things that are convenient to measure. Take the example of CO_2 emissions, whose components are measured with different degrees of precision. There are established accounting protocols and a clear unit with tons of CO_2 equivalents readily available to anyone interested in measuring CO_2. If firms provide the figure (usually voluntarily), we know a lot about a firm's emissions in the past. But we do not yet know the firm's future emissions, which are based on the decisions that the firm is making today. We also know very little about emissions in the firm's supply chain, usually referred to as scope 3 emissions.[1] Scope 3 emissions need to be estimated even by the firms who are disclosing them.

Furthermore, the reported CO_2 data needs to be put in context to truly understand a firm's impact on society. Let's have a little thought experiment to illustrate the concept of additionality regarding the CO_2 emissions of a firm. Assume there is a small town in Oregon that consumes 100 percent of the electric-

[1] See the EPA for the definitions: https://www.epa.gov/climateleadership/scope-3-inventory-guidance.

ity from a small hydro plant producing 1 megawatt. Assume the plant is at capacity – meaning that it cannot produce a single additional kilowatt. Now, Amazon decides to put a massive AWS server nearby, and that the energy demand for the warehouse is also 1 megawatt. Amazon signs a contract with the hydro plant to purchase 100 percent of their electricity at a premium. The plant decides to sell to Amazon, and now the town is forced to buy electricity from the grid. The problem is that the electricity is from a coal plant, which clearly will produce CO_2 emissions. The question is, who is the one responsible for producing the CO_2? According to the expenses, Amazon is purchasing clean energy producing zero emissions and the town is buying the dirty energy. Hence, the Additional impact of Amazon is 1MW of dirty energy. The actual economic accounting should assign 100 percent clean to the town, and 1MW dirty energy to Amazon regardless of the expenditure shares. However, in current accounting practices and in what Amazon would report to Carbon Disclosure Project, there would be zero emissions attached to this particular warehouse. Solving the problem of accounting for additionality has proven to be one of the most difficult tasks.

Is it easy to measure the treatment of female employees? Of course not! The share of women on the board is likely a very coarse indicator for discrimination. But we should still try to assess firms in terms of how they are handling such an important issue.

Understanding the limitations of the measurements is crucial. As we said, if something is important to society, it should be measured, but it also should be understood and recognized that the measurement is imperfect. This is particularly crucial for regulators to understand. For example, a complex problem such as discrimination and mistreatment of historically disadvantaged groups in the labor force cannot be summarized by simply looking at the proportion of these groups in management. If regulators focused on this statistic, firms might comply and achieve the right proportion of these groups in management but continue to mistreat them. In other words, firms might hit the target, but miss the point.

The point is, and should be, about treatment of historically disadvantaged groups regardless how many are in the organization. There is no possibility that the perfect measurement of the intentions can be achieved, so we need to learn to live with imperfect measurement – not only for discrimination but for almost all social aspects. This is a delicate balance that is difficult to navigate. On the one hand, if an issue is important, it should be measured – regardless how hard or uncertain the measurement is. On the other hand, what is done with the measurement is a matter of understanding its precision and accuracy.

As we showed above, the notion that CO_2 is properly measured, as suggested by many academics and practitioners, is problematic. In the last two decades, firms have been increasingly willing, to disclose their CO_2 emissions. The reported emissions have been collected by the Carbon Disclosure Project. Participation is voluntary, and the verification of that data is also voluntary. This implies that there are many missing observations. Many firms choose not to report and those that report are not necessarily representative of all firms. This is particularly pervasive in the carbon market. In total, 80 percent of the scope 1 and 2 CO_2 emissions provided by TruCost have been imputed, and about 95 percent of the scope 3 is imputed.[2] This procedure makes sense if we are interested in obtaining an estimate of the "world" CO_2 emissions; but should these imputations be used for regulatory purposes? Is this truly a better measurement than the number of historically disadvantaged employees in management?

AGGREGATING DIFFERENT ISSUES

Should we standardize what ESG issues rating agencies should take into account and how important they are? The short answer is again no.

Let us assume that problems around the measurement of different issues, such as discrimination and climate change risks, have been resolved. The next question is if we can put these issues together in a single score – as is customary for the ESG rating industry. In other words, can we settle on one aggregation rule for an overall encompassing ESG rating? No chance!

Aggregation is fundamentally about preferences, and individuals have different preferences. Some people will think climate change is the most important issue, others feel more passionate about discrimination, others about biodiversity, and others about poverty. How can a single score capture the heterogeneity in preferences? Who are we to tell anyone what is important to them?

The standardization of ESG ratings entails the existence of what is known as a social welfare function. This is a function that, as its name indicates, captures what the preferences of society are. The social choice discipline in economics (and mathematics) has several interesting results regarding this function. First, when there are more than two issues, and preferences are heterogeneous, it is impossible to guarantee that the social welfare function exists. Unless it coincides with the preferences of a single person – which Arrow denominated the "benevolent dictator." Indeed, in the 18th century, the Marquis de Condorcet proposed a paradox in which three rational individuals will behave irrationally when pairwise comparisons are made. Assume that there are three fruits: Apples, Bananas, and Coconuts. One agent prefers Apples to Bananas to Coconuts; the second one prefers Bananas to Coco-

[2] Trucost is a product of S&P Global assessing risks relating to climate change, natural resource constraints, and broader environmental, social, and governance factors. It is widely used to measure CO_2 emissions.

nuts to Apples; and the third one prefers Coconuts to Apples to Bananas. Each agent is individually rational and let us assume we name them members of our Congress to make the decision about which fruit we should serve. Humans tend to make pairwise comparisons (A versus B). Assume we vote, and each agent gets a vote. When comparing Apples to Bananas, Apple gets two votes (agents 1 and 3), and Banana one (agent 2). So, Apples are better than Bananas. When we compare Bananas to Coconuts, the first one gets two votes (agents 1 and 2) and Coconuts gets one vote (agent 3). So, Bananas are better than Coconuts. We would assume that if we were to compare Apples to Coconuts, it should be the case that Apples are better. However, that is not the case. Apples would get one vote (agent 1), and Coconuts would get two votes (agents 2 and 3). Hence, even though we can represent the preferences of each individual, we often cannot represent the preferences of the aggregate.

Some argue that the default objective should be financial materiality and the maximization of stock returns. This might indeed define an aggregate index and it might make sense when you sell your data to investors, but it is a poor social welfare function when thinking about resolving the underlying issue (Simpson, Rathi and Kishan 2021). For example, assume that child labor is not materially important in a sector. Should we not measure it? Should we measure it and not include it in the index? This is a very difficult problem to solve, especially because almost no one has asked investors and consumers about their preferences. Therefore, not surprisingly, the rating agencies are proposing different aggregation rules – which generates another source of discrepancy.

In addition, ESG rating agencies currently use quasi-linear aggregation rules. Our research shows that this implies certain trade-offs that would make most people feel very uncomfortable. Again, a simple example makes the point. Assume that you measure discrimination against two different groups, women and LGBTQ+. The aggregate score of each firm is determined by the average between the scores for women and for LGBTQ+ (the example holds for any linear weighted average technique). Imagine that one firm gets a score of 60 for discrimination of women and 20 for people identifying as LGBTQ+, the aggregate score being 40. Imagine the firm feels bad because they think they are discriminating LGBTQ+ too much (score of 20), and they come to the rating agency and ask: "I want to keep my overall score constant, is it okay if I discriminate woman more and a little bit less LGBTQ+ such that the aggregate remains at 40?"

Most would say that the question is unacceptable, and that the individual should improve the treatment of women without deteriorating the treatment of any other group. In fact, treating one individual correctly is not a license to treat another one badly. This notion, however, is not captured by a linear aggregation rule, but by a non-linear one. Our research shows that the ESG scores can indeed be approximated quite precisely with a linear aggregation rule. This means that firms can make decisions that imply trade-offs that could be unacceptable to most citizens. Therefore, it is possible to compute how many tons of CO_2 a company can emit more if it adds one more woman to the board - keeping the overall score constant

Some people are actually willing to trade-off between certain issues but most likely not in a linear way. For instance, some could be willing to accept a small deterioration of the human rights record of a firm if this is accompanied by a massive reduction in CO_2 emissions. But this calls for more research on preferences and aggregation functions. Hence, a standardization of ESG ratings would also disincentivize improvements about how to build the optimal aggregation function for a given investor.

CONCLUSIONS AND POLICY IMPLICATIONS

Would the world be a better place without ESG ratings? Our research implies no. The information that ESG raters produce is valuable. And assessing ESG performance is not only conceptually challenging, but also labor intensive. ESG rating agencies need to assess many issues: CO_2 emissions, water, biodiversity, labor treatment, discrimination, inclusion, product safety, marketing practices, supply chain, lobbying, corruption, and taxes, among others. They need to make this assessment for thousands of companies and update it regularly. If you find this task daunting, you are in good company. Of course, it is costly to undertake, but it is worth doing, because ESG issues matter. Rating agencies can harness economies of scale, and competition among them helps to drive down costs, if the market is set up the right way.

What would be a good market setup for ESG ratings? The key is to create a competitive market, where competition is centered around the quality of measurement. We believe there are three useful steps regulators should take: standardize ESG disclosure (not the ratings), enhance transparency about methodologies, and encourage compatibility between rating systems.

When it comes to standardization, regulators need to distinguish between firm disclosure of ESG data and ESG rating agencies themselves. Firms often rely on disclosure frameworks such as the Global Reporting Initiative, the Sustainability Accounting Standards Board, and the Greenhouse Gas Protocol to publish ESG-related data. For instance, if firms count CO_2 emissions differently, it would be hard to interpret that data. Hence, standardization with the help of disclosure frameworks is useful. ESG rating agencies can then use this data, check if it is credible, add data from third-party sources, and thus form an opinion about the ESG practices of the underlying firm. If valid and standardized ESG data is widely available, ESG ratings can compete more on interpreting the data,

and less on collecting the data privately. There will still be divergence, but it will be divergence in opinion, not disagreement about facts.

With regard to ESG ratings, we believe standardization of how and what ESG ratings measure, with the aim of making them diverge less, would ultimately result in less reliable information. However, regulators should increase transparency about measurement practices and aggregation rules. Indeed, without transparency, there cannot be any competition between the best measurement practices or aggregation rules.

Finally, regulators (but perhaps also market participants themselves) should develop a taxonomy for how the issues within E, S, and G are broken down. There are many reasonable ways to slice and dice ESG issues, but the fact that each rater does it differently makes comparison across raters unnecessarily difficult. This is a compatibility problem, similar to the problem of when you switch your cellphone, you cannot use your old charger anymore, which provides exactly the same function, just with a different plug. From the perspective of the users of the ESG ratings, it is far more convenient if the sub-scores are available in the same set of categories. This makes it easier to compare and switch to alternative providers, which fosters competition.

In sum, can the data and procedures be improved? Yes. Can the discrepancy be made smaller? Of course. But does that mean that the data today is useless, that it should not be used as a measuring stick, or that some of it needs to be standardized or even disregarded? No. ESG ratings are useful and relevant today, and it is essential to maintain investment and innovation in ESG ratings. The existing shortcomings are not a reason to resign. Instead, they call for redesign.

REFERENCES

Berg, F., J. F. Kölbel and R. Rigobon (2022), "Aggregate Confusion: The Divergence of ESG Ratings", *Review of Finance* 26 (6), 1315–1344, https://doi.org/10.1093/rof/rfac033.

Simpson, C., A. Rathi and S. Kishan (2021), "The ESG Mirage", *Bloomberg*, 10 December, https://www.bloomberg.com/graphics/2021-what-is-esg-investing-msci-ratings-focus-on-corporate-bottom-line/.

The Economist (2022), "The Signal and the Noise", Special Report, 23 July, https://www.economist.com/special-report/2022/07/21/the-signal-and-the-noise.