ESG Consensus Ratings – The Key to Asset Owner Oversight

Utilizing Artificial Intelligence and Machine Learning to Overcome the Subjectivity of ESG Rating Vendors[1]

By Steve Glass - Co-CEO, Abel Noser Holdings

Introduction

The centrality of ESG-related issues to investors is borne out by the rapid integration and growth of ESG-oriented retail and institutional investing. In this regard, ESG-related initiatives are ultimately about managing risk. For example, as noted by the Organization for Economic Co-operation and Development (OECD), a poor environmental record may make a firm vulnerable to legal or regulatory fines/sanctions; socially, mistreatment of workers and dissatisfied employees may lead to higher absenteeism, lower productivity and weaker client servicing/relationships; and weak corporate governance may incentivize and/or enable unethical behaviors related to pay, accounting irregularities and even fraud.[2] For all these reasons, identifying and addressing the various ESG issues germane to a corporation (i.e., "material"), is a quintessential exercise in risk management – for company management, investment managers thinking about holding that security in their investment portfolio, and asset owners concerned whether the manager is acting in accordance with fund policies.

However, even for practitioners who care deeply about ESG, a key obstacle they face is that almost all ESG-related data consists of raw metrics. Consequently, the user must have the internal resources to both collect and aggregate the data, as well as sufficient internal subject-matter expertise to consume the raw data and assess whether follow up is warranted. The one exception is ESG ratings.

ESG ratings purport to assess and rank the degree to which each company manages its respective ESG risks. Today, several vendors offer services that assess company compatibility with ESG-related ideals and provide associated ESG ratings/scores. In theory, practitioners thereby need not consume the raw data themselves. Rather, they can utilize these ratings/scores to determine which companies are best managing their respective material ESG risks; and focus their analysis on monitoring company ESG ratings/scores.

In this regard, it is easy to see how ESG ratings/scores can have applicability at every stage of the investment cycle: asset allocation, investment universes, portfolio construction, investment selection, risk management, and regulatory/client reporting. And ideally, a better ESG rating will correlate over time to higher returns.

Unfortunately, as discussed below, the firms that provide ESG ratings/scores often yield very different rating assessments. Indeed, it's not uncommon for a company to be rated very highly by one rating provider while simultaneously being rated very poorly by another. This lack of consensus has historically represented a significant challenge for practitioners hoping to develop policies and investment decisions based on those ratings/scores. Happily, recent developments in the ESG ecosystem offer a solution to this problem.

[1] *Acknowledgments: Benjamin Webster, CEO and Dale Neibert, Managing Director of OWL ESG Inc. played a central role in formulating the positions and many of the references cited herein. This article could not have been completed without those efforts.*

[2] OECD, ESG Investing: Practices, Progress, and Challenges, (2020)

This article is divided into two sections. The first section explains in greater detail why individual providers of ESG ratings/scores, even with the best of intentions, often assess the same company very differently. The second section describes how artificial intelligence and machine learning have helped provide an elegant quantitative solution to this problem.

The Challenge - Non-Standardized Ratings

Currently, almost all ESG-related data disclosures are subjective, non-standard, self-reported, and unregulated. Many companies don't even collect, let alone disclose, all the relevant data. This presents severe challenges to ESG rating vendors.

To be clear, rating vendors typically employ armies of analysts that focus on assessing the ESG risks and opportunities of companies. They evaluate hundreds if not thousands of data sources to glean ESG information about each company they rate. Then they take that information and feed it into their databases and models. The result is a robust database of scores and rankings produced by each respective rating vendor.

Unfortunately, because of the disparate and inconsistent nature of company data disclosures, every ESG data provider, by definition, must make important decisions about which data sources to use, how to weight various factors, and then apply its own ethical judgements and algorithms as to the key considerations associated with each respective industry. No surprise, inconsistency is the norm.

The OECD report, for example, observed that the methodologies used by the major providers of ESG ratings/scores, were all "quite different."[3] These differences were, of course, subsequently manifested in the ratings/scores those firms assigned to each company. In support of this view, the OECD also cited a 2019 analysis conducted by State Street Global Advisors, which, as shown in Table 1 below, demonstrated that the correlation of the ratings issued by these firms ranged as low as 0.47 and never exceeded 0.76.

Table 1

Table 4. SSGA Assessment of R^2 of ESG ratings among major score providers

	Sustainalytics	MSCI	RobecoSAM	Bloomberg ESG
Sustainalytics	1	.53	.76	.66
MSCI		1	.48	.47
RobecoSAM			1	.68
Bloomberg ESG				1

Source: State Street Global Advisors (2019)

More recently, a 2021 study conducted at MIT found that while the correlation of bond rating agencies Moody's and S&P was 0.92, the correlation of ESG rating agencies ranged from only 0.38 to 0.71.[4]

To illustrate the practical challenges such low correlations present to investors, the OECD compared the ratings provided by Bloomberg, MSCI, and Refinitive to every constituent of the S&P500 and STOXX600 indices. As one would expect, the analysis highlighted wide differences between the three rating vendors, with the average

[3] The firms reviewed were Bloomberg, Thomson Reuters, FTSE, MSCI, and Sustainalytics.
[4] Berg, Koelbel, Pavlova, Rigobon, ESG Confusion and Stock Returns: Tackling the Problem of Noise, Nov. 19, 2021

R^2 for the S&P500 being only 0.21, and the STOXX600 being even worse at 0.18. Figures 1 and 2 below show the distribution of ratings for the two index constituents.

Figure 1 – S&P500

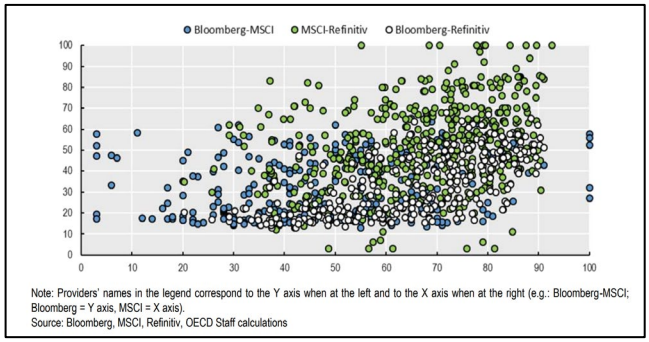
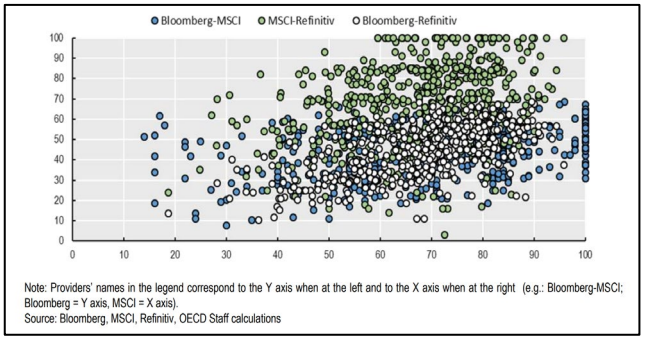


Figure 2 – STOXX600



The OECD identified a variety of causes for the different weightings generated by each firm. These included:

- Differing ESG frameworks
- Different definitions of 'materiality'
- The inclusion (or exclusion) of news controversies
- Different KPIs within the E, S, and G pillars
- Different sub-category indicators
- Subjective weighting judgements

Similar to the OECD studies, the ESG research and ratings firm OWL ESG Inc. conducted extensive research in 2018 on the correlation of 14 individual ESG rating vendors. OWL ESG's study found that, of the top 100 US companies rated by those firms, only 12 companies were represented in all 14 lists. In addition, when measuring these 12 companies by rank, there was an average dispersion of 25 ranks.[5] These inconsistencies were a function of three key causes:

Different ESG Factors
Each ESG rating firm, based on their diverse backgrounds and focus, utilized what they considered to be the most relevant ESG factors for each respective industry and sector. Perhaps not surprisingly, in the study there was only about a 50% overlap between the factors that any two rating firms viewed as material (in producing their ESG scores for any given company).

Different Weighting of ESG Factors
Based on each firm's particular expertise and knowledge, even when the firms identified similar ESG factors, they weighted those factors differently in their algorithms. Furthermore, the firms also combined and calibrated those ESG factors differently. These subjective weighting choices obviously created further divergence between each company's ESG scores and ranks.

Different Data Sources
Without a single consolidated data source, as a practical matter, each rating vendor gathered data from a variety of news outlets, non-government organizations (NGOs), research firms, and other sources. Some data sources were, of course, used by multiple vendors, but in many instances each vendor relied on unique and different sources as well. Further, even when the data sources were the

[5] *OWL ESG Methodology*, page 1 (2018)

same, the rating firms often looked at different time periods. It followed that, even had the algorithms been the same (which as noted above, they were not), by inserting these differing data sets into their models, the rating firms inevitably produced different results.

It's important to note that the historical faults in ESG rating practices observed by the OECD, MIT, and OWL ESG studies are inherent symptoms of the industry. Unless and until disclosures become uniform (whether by common practice or regulatory fiat), rating providers will always need to make choices. And where there is choice, there is subjectivity.

Interestingly, while the ESG rating of a company may vary widely between one rating firm and another, the bond rating credit scores of those same companies are much less divergent. As noted by the OECD, this implies that good (or bad) ESG ratings are largely dependent on the methodologies of the rating vendors, not the underlying financial strength of that company. More to the point, investors cannot have confidence that the existence of a good (or bad) ESG rating from any individual rating provider correlates to future enhanced returns of that company, or truly aligns with a particular societal value.

Further diminishing the merits of using a single individual rating vendor is the risk that said firm might award their ESG ratings based, in part, on conflicts of interest (that result in the utilization of non-ESG considerations). For example, another MIT study published in August 2021 documented practices by Refinitive ESG in which they retroactively altered their historical ESG ratings in a manner that improved the correlation of those revised ratings to stronger investment performance. [6] Regardless of the firm's motive for doing so, this practice gave a false impression that had managers used Refinitive's ratings to construct portfolios, they would have achieved stronger returns. The study noted that one potential reason for this practice was to attract new subscribers to their ESG rating service.

Another example of non-ESG considerations factoring into the calculation of an ESG rating/score, is the awarding of better ESG ratings to affiliated companies (or companies with common shareholders). This was the focus of a recent 2022 academic study that reviewed the ratings of the ESG rating and research firm KLD Research & Analytics. This study quantified that KLD's ratings of MSCI affiliated companies significantly improved *after* KLD was acquired by MSCI.[7]

To be clear, ESG ratings can be a valuable (and maybe essential) tool for translating raw data into a more sophisticated quantitative metric that can be used to build portfolios and/or exercise prudent oversight. And diverse methodologies and approaches certainly help enrich the information available to investors (a fact recognized and applauded by the OECD). However, most practitioners don't have the resources to subscribe to all these data sources nor distill such quantities of data into actionable information.

Consequently, the existence of those differences reduce the usefulness of the assessments of any single ESG rating firm (a fact also recognized by the OECD). As a practical matter, the large discrepancies seen across multiple rating providers (whether due to subjectivity or conflicts of interest), implies that the shape of each investor's security selections could be driven by the choice of rating provider rather than the underlying merits of the actual securities.

[6] Berg, Fabisik, Sautner, Is History Repeating Itself? The (Un)predictable Past of ESG Ratings, August 24, 2021
[7] Tang, Yan, Yao, The Determinants of ESG Ratings: Rater Ownership Matters, (June 6, 2022)

Bottom-line, without a means to reduce those rating differences, the value of ESG ratings/scores are themselves thereby reduced. More to the point, for investment practitioners looking to use ESG scores to manage portfolios, design financial products, or monitor those firms that do, such a wide range of ratings can be disconcerting. Rhetorically speaking, practitioners who want to assess adherence to ESG standards may be forgiven for asking how they can rely on ESG ratings when there's such a wide divergence of opinion even among the experts. Similarly, skeptics of ESG may worry that ratings might provide a false sense of security.

The Solution – Consensus Ratings

A very practical response (until greater disclosure standardization occurs) is to leverage technology in a manner that consumes ESG data from multiple rating vendors (as well as other sources) and optimize that data in a way that reduces the inherent subjectivity between any two rating providers - in effect creating a 'consensus rating/score.'

Each ESG rating agency likely spent years studying ESG in general, examining which ESG factors had a material impact on each industry and sector, and optimizing their algorithms. They had good reasons for why they chose which ESG factors to use in their ratings processes. The use of statistical models that optimize these inputs to generate consensus ratings/scores, in essence, leverages the collective efforts, knowledge, and expertise of all the rating vendors, while at the same time managing the subjective choices as to which ESG factors are important to which industries.

As one might imagine, this approach results in aggregated ESG scores/metrics that necessarily incorporate significantly more company-specific ESG data than any single rating vendor. For example, one firm that has taken this approach is OWL ESG Inc. OWL ESG consumes data from over 500 sources, among them 14 well known ESG rating and research firms (both generalists and specialists), news & media outlets, NGOs, government databases, unions, activist groups, and other various public sources.[8]

In aggregate, OWL ESG collects over 100 million data points each quarter and rates every company on approximately twice as many industry-specific ESG factors than any single provider. The effort to distill this amount of data into a coherent and intellectually rigorous consensus rating is not trivial. OWL ESG, for example, applies a four-step process:

- Identification of common ESG factors

[8] Brief descriptions of these types of sources include the following: **Generalist ESG Research Providers** are firms that typically conduct extensive fundamental research to gather, synthesize, and analyze environmental, social, and governance data about companies. They also aggregate the data to create scores and metrics that can be used to gauge the performance of companies compared to peers regarding ESG factors. OWL ESG consumes data from many generalist ESG research providers. **Specialist ESG Research Providers** are firms that perform much of the same work as the generalists, but instead focus heavily on usually one of environmental, social, or governance ESG data and metrics. OWL ESG consumes data from a number of specialist ESG research providers. **Controversy ESG Research Providers** are firms that focus their research on news and other media sources to find controversies at companies involving ESG factors in an attempt to gauge the severity of ESG-related controversies. Many of the generalists have controversy products as well. OWL ESG consumes data from a number of controversy ESG research providers. **Public Sources** or public source data providers come in many shapes and forms. Some are founts of deep research like the Carbon Disclosure Project. Others are public source ratings providers like Glassdoor. Some are government sources like the Environmental Protection Agency. While still others are questionnaires from international organizations like the UN Global Compact. OWL currently collects and processes data from 400+ public sources.

- Data conversion and normalization
- Data aggregation and optimization
- Formation of peer universe comparisons/rankings

Taking each of these steps in turn.

Step 1 - Identification of Common ESG Factors

The first step is to identify which ESG factors are important (i.e., material) to each industry and/or sector according to the leading ESG rating and research organizations; and if possible, discern any common views among their recommendations.

This step includes examining how each rating provider organizes their data at the industry and sector level, how they separate different data elements (e.g., females on the board of directors) into subthemes (e.g., board diversity), and how they then aggregate those subthemes into higher-level ESG factors (e.g., diversity). Lastly, this step reviews how each rating vendor combines those ESG factors into the E, S, and G pillar scores, which in turn, comprise their overall ESG ratings.

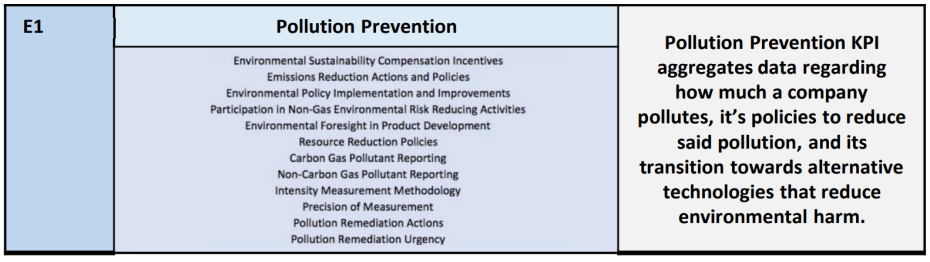
Based upon the aggregated consensus data, OWL ESG for example, identified 12 themes which they characterized as key performance indicators (KPIs). Those 12 KPIs roll up into one of the three pillars of ESG. Figure 3 below displays the 12 KPIs and their respective environmental, social, and governance pillars.

Figure 3

Environmental	Social		Governance
	Employer	Citizenship	
Pollution Prevention	Compensation & Satisfaction	Community & Charity	Board Effectiveness
Environmental Transparency	Diversity & Rights	Human Rights	Management Ethics
Resource Efficiency	Education & Work Conditions	Sustainability Integration	Disclosure & Accountability

Importantly, the 12 KPIs collectively are based on thousands of ESG data elements that are combined into over a hundred subthemes (which often have different variations depending on the industry). For example, Figure 4 below shows the key subthemes that roll up into the overall assessment of a company's ability to manage its pollution prevention KPI (which is one of the three KPIs that comprise the environmental pillar).

Figure 4



Step 2 - Data Conversion and Normalization

Step two involves intense data conversion - designed to feed data elements into the step one KPI schemas - to produce ESG consensus metrics. As noted earlier, the actual data elements come from multiple ESG data sources and in many different formats. That data is sometimes expressed as raw data (e.g., the amount of carbon emissions) and other times as a simple yes or no. In fact, in some instances the data may already be expressed as a rating. Ultimately, the conversion process takes those raw inputs and converts them into scores between 0 and 100.

In essence, for each data source, a distribution curve of results is generated for every company covered by that data source. This analysis is then repeated for each company on an iterative basis based on the data received from each and every data source available.

Once the analysis has been conducted on every company and for every data source, the results are then normalized. For example, if five data sources report on carbon emissions, Source 1 might have 60% of companies reporting low carbon emissions, such that a company with typically low emissions receives a score of 40. In contrast, Sources 2–5 all may have approximately 52% of companies reporting low emissions, such that companies with typically low emissions receive a score of 48.

To bring Source 1 in line with the other sources, the carbon emissions scores for the companies evaluated by Source 1 would therefore be adjusted, such that a company with typically low emissions would receive a score closer to 48 instead of the score of 40 originally published by Source 1. In this manner, there would be a normalized baseline of scores between 0 and 100 for all companies reporting a similar carbon emissions level.

However, the fact would remain that any given company might still receive five different 0–100 carbon emissions scores, given that each of the five rating firms used different raw data for their assessments. This then brings us to the third step – the data aggregation and optimization (i.e., weighting) phase.

Step 3 - Data Aggregation and Optimization

The goal of the data aggregation and optimization phase is to build a general consensus from the disparate data sources. In this phase, the data elements that were all scored on the 0–100 scale are then aggregated up to their appropriate KPIs. At the KPI level, various rules pertaining to the weights given to each factor and metric, whether there should be a data threshold, and the timeliness/staleness of each data point, are applied.[9] More specifically, it addresses the challenges associated with rating vendors using different data sets and time periods.

This weighting process recognizes that each of the rating vendors and other data sources, have spent significant resources optimizing their models to include ESG data that *they determined fit their definitions of "material"*. Consequently, the weighting approach implicitly assigns a higher weight to data associated with the ESG factors that more sources felt was materially relevant for a given company within its industry.

[9] For example, if more sources are reporting on certain ESG factors mapped to a KPI (e.g., carbon emissions) than other factors mapped to that same KPI (e.g., non-carbon emissions), those reported factors will automatically have more "weight" within that KPI's score.

Of course, the mechanics of weighting ESG data is easier said than done. First off, a minimum amount of ESG-related information on a company must be available to receive a score for a KPI. In this regard, it's important to remember that a large portion of the ESG data ultimately comes from the companies themselves; and ESG disclosure is a voluntary process. No two companies disclose information on the exact same ESG factors.[10] This can make it very difficult to create valid comparisons between companies, even for two companies in the same industry.

Additionally, even when sufficient data exists for a particular KPI, there may be instances where a certain data point from one source diverges substantially from the data points provided by the other sources. Notwithstanding such divergence, if the source of that data point is deemed credible, it may still be included in the calculation of the subtheme score.

In these instances, the consensus rating/score algorithm determines the credibility of a source by looking at all the data on all the companies, which that source has published since inception. The algorithm then compares that data to the data published on the same company by all other sources. The higher the correlation between a source's historical data and that of all other sources, the more credible that source is deemed to be with respect to the subject data-outlier. In a similar fashion, the OWL ESG algorithm weighs more recent data higher than older data.

The net result of the data aggregation and optimization phase is the generation of intellectually sound, consensus driven KPI, ESG pillars, and overall ESG ratings/scores.

Step 4 - Formation of Peer Group Universe Comparisons/Rankings

Once the data aggregation and weighting are complete, the fourth and final step is possible. This is the creation of relevant peer groups and the associated ranking of each company within those groups. OWL ESG, for example, segments companies into peer groups based on geographic regions, and within each region the company's sector, sub-sector, and industry. The purpose of dividing companies into peer groups is to better isolate the most relevant variables affecting the financial performance of companies within those peer groups.

Moreover, it's important to note that not all ESG factors are as critical to some industries and sectors as they are to others. In fact, some ESG factors have been shown to have a positive material effect for some industries while simultaneously having a negative material impact in other industries.

By establishing relevant peer group universes, each company can be assigned a monthly rank and percentile within their appropriate peer group for every KPI, ESG pillar, and overall ESG rank. This enables investment practitioners to quickly identify, with greater confidence than ever before, how companies compare to their peers across all ESG metrics.

Conclusion

To sum up the practical challenges currently facing ESG ratings, inconsistent disclosure practices lead to inconsistent data that's available to rating vendors. Consequently, each rating vendor applies its own

[10] There can be many reasons a company may not disclose specific ESG information, ranging from the company having "bad" metrics to not having a process in place to track that information.

proprietary subjective algorithms and weightings to that data (in terms of what is materially important to each respective company). This in turn results in inconsistent assessments (i.e., ESG ratings) of how well each company is managing their respective ESG risks.

The solution to this problem is the generation of ESG consensus ratings/scores. Once constructed, in our view, the consensus rating/score approach enjoys a number of advantages over the use of any single ESG rating firm.

Significantly greater inputs: The consensus rating/score approach, by definition, consumes multiples of data more than any single rating vendor. For example, OWL ESG consumes over 100 million data elements from over 500 sources, including 14 well known ESG research firms, news and media outlets, non-government organizations, government databases, unions, and activist groups.

Reduced subjectivity: The consensus rating/score approach essentially leverages the "wisdom of the crowd" theory, by which ratings are derived from hundreds of inputs and reflect a weighted score based on the number of ESG rating providers that view each respective E, S, and G metric as relevant for a given industry. The resulting statistical optimization reduces bias and error and generates a consensus viewpoint for every company covered. This shifts the derived ESG ratings from being heavily based on a single provider's subjective judgements/priorities toward a more quantitative statistical optimization of the marketplace's broad consensus.

Significantly more coverage: Because OWL ESG receives data from so many different sources on so many different companies, it has the largest ESG data set in the industry, generating coverage on over 31,000 companies worldwide. This is significantly broader coverage than other providers who typically rate 8,000 - 20,000 companies.

Monthly frequent scoring: Similarly, because more data is consumed, specific data that warrants a change in rating is uncovered quicker. This enables ESG ratings/scores to be updated monthly (instead of annually) leading to more dynamic metrics appropriate for portfolio management, indexing, and asset owner oversight.

An additional, very practical, benefit of monthly-refreshed data is that annual company CSR reports (in which each company discloses their ESG updates) are often published on different timetables than the ESG rating firms' annual updates. This lack of coordination can sometimes result in over a year's delay before ESG ratings reflect the new company data. Monthly ESG rating updates avoid this risk, and thereby provide users with an information-edge by consuming company CSR reports in a timely fashion.

Bottom-line, ESG consensus ratings/scores provide greater transparency and increased confidence for investment practitioners who wish to develop high-level views on such issues as:

- Validating (or challenging) the achievement of ESG ideals and manager representations
- Putting company ESG ratings into context by utilizing comparisons to relevant industry peer groups
- Tracking positive (or negative) ESG trends
- Quantifying and unbundling ESG attribution (e.g., "*What if E conflicts with S or G?*")

- Flagging ESG outliers
- Complying with ESG regulatory frameworks and client policies/directives

A crowd's collective intelligence [when independent and diverse] will produce better outcomes than a small group of experts…

James Surowiechi, The Wisdom of the Crowd, 2004