ESG Ratings and Rankings

All over the Map. What Does it Mean?

Dr. Jim Hawley
Head of Applied Research
TruValue Labs

Table of Contents

TRUVALUE LABS

Introduction and Scope of the Problem

It has long been noted—usually with significant concern—that ESG/SRI rating and ranking organizations (and indices as well) frequently rate the same corporation's E, S and G elements differently.[1] This is true for those organizations that give overall ratings, as well as those that provide far more granular ratings to specific E and S and G content (e.g. supply chain, board structure and performance, even carbon use and mitigation). As one study noted in regard to company CSR (corporate social responsibility, increasingly the corporate flip side of investor-oriented 'ESG') reports, they are 'selective, subjective and not comparable'.[2] In much of the academic literature, CSR is often the object of study rather than ESG. The same can be said for ESG raters and rankers. There is an increasing academic and practitioner literature on ESG/sustainability raters, e.g. MSCI (KLD), Sustainalytics, Asset4,[3] Vigeo Eiris.[4] For the purposes of this paper I use the terms ESG, sustainability, CSR and CSP (corporate social performance) interchangeably.

The academic CSR literature extends back many decades as the same evaluation and rating problems were evident in the decades-long history of CSR and CSP.[5] The purpose of this short paper is to survey some of the problems and concerns that raters and rankers present for the analysts, scholars and users of this information, to present an overview of what I think is the best of this literature, and in conclusion to suggest what may new directions for change.[6]

In some regards the ESG rating and ranking compatibility problem stands in contrast to the high level of correlations among and between credit rating agencies of corporate default probabilities.[7] Yet a closer approximation of the ESG ratings general lack of correlation is found in the disagreements among credit rating agencies in regard to areas such as corporate governance and in investment prohibitions and capital requirements among institutional investors.[8] It is likely that in the default case the data is harder, more timely and standardized, while in the latter cases, it is softer. A better parallel is between ESG raters and rankers and sell-side stock analysts who typically have significantly varying buy, hold, sell recommendations, despite theoretically having access to similar if not identical information. This is also true of pure quant trading, with algorithms substituting for human calls.

In the title I suggest that the various ratings and rankings are all over the map. There are upwards of 600 'products' from over 150 organizations providing ESG data and dozens of others which rank firms along a host of dimensions, e.g. Fortune and Newsweek.[9] In spite of the large number of data providers, raters, and rankers, the industry is rapidly consolidating, with MSCI and Sustainalytics the current leading players. This in no way implies that the problems of quality have changed.

One simple example: a comparison of KLD (MSCI) against Fortune magazine's 'Best 100 Firms' looks like this, see Figure 1. It has correlation of only r=14%[10].

Figure 1. Actual vs. Perceived CSP
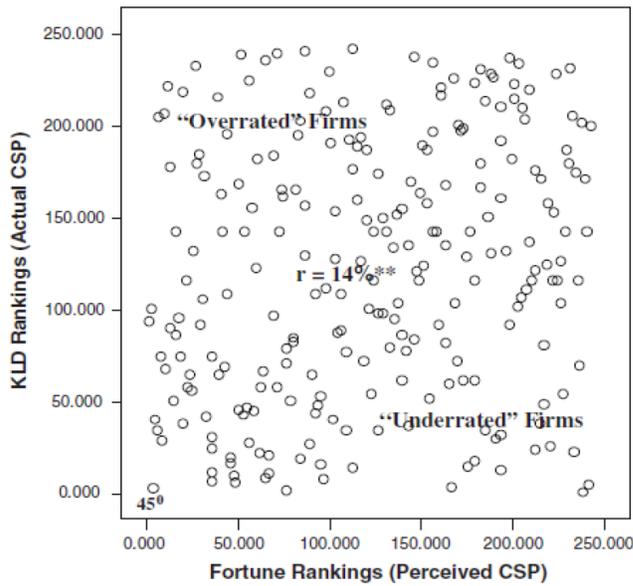
The point here is not the KLD (MSCI) is the North Star of ESG data, nor that Fortune's Best 100 are not 'good' or worthy, but rather there is no way to know what to make of such a low correlation, at least without digging deeply into data sources and methods, to the degree that is possible (and far too often it is not possible, or certainly not easily done).

Another more specific example is the huge divergence of evaluation in the case of ethical palm oil sourcing (an S and E issue) when considering the inter-rater reliability of 30 companies by three evaluators.[11] Again, we see little correlation, even on this granular level.[12]
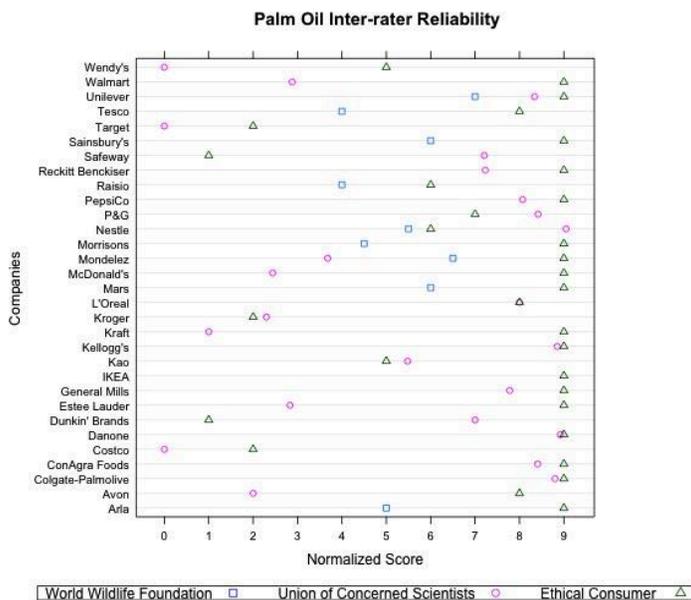
The takeaway in short: Good analysts do not take third party raters and rankers simply at their word, they need to dig deeply into E and S and G data and methods just as they do into standardized financial data, which itself often has significant wiggle room and does not necessarily speak for itself.[13]

Both data and methods are formidable hurdles in the ESG/CSR/CSP arenas, as most providers of ratings and formulators of rankings are not transparent, or only transparent to a degree. Windolph summarizes six of generic problems noted by researchers.[14]

- **Lack of standardization**. There is a diversity of approaches, hence of results, little evaluation of the multiplicity of approaches, and no comparability. Comparability can be developed to some extent by elaborate mapping exercises, which many researchers do, but these inevitably involve inaccuracies and some guesswork.

- **Lack of transparency.** There is rarely full disclosure of methodology, criteria, or threshold values and levels.
  - **Biases**. Biases along several dimensions exist.
    - **Geographical bias.** As ESG is more developed in Europe, there is a tendency to use these standards.
    - **Factor bias.** In some cases, there is no transparency regarding weighting of various categories or indeed the categories themselves. E.g. some firms use economic as well as ESG factors; others focus mainly in their versions of what is 'ethical', and may minimize, for example, environmental factors
    - **Selection bias.** Additionally, there can be bias towards investors or, alternatively, stakeholders. There is bias towards l rating larger firms more positively as these firms report on ESG factors in greater proportion and in greater detail than smaller firms. This is sometimes called "check the box bias."

- **Trade-off problems.** Some raters focus on a single, top level score; others are very granular. There can be a tendency to add apples and oranges. For example, Wells Fargo & Company has historically done very well on diversity, including on its board. Yet as recent scandals reveal, the board and top leadership was flawed in part due to their long tenure and insular nature. Thus, diversity appeared strong (as a G factor), yet another G factor, tenure, was overly long compared to industry peers.

- **Lack of credible information**. This has a number of elements. Some raters use questionnaires and interviews with firms, minimizing independent information. Others rely heavily on firms' sustainability reports. Some firms use only independent information. Few firms are transparent about their sources, the weighting of these sources and how, if not using pre-set

mechanical weighting, interpretation plays a role.

- *Lack of independence*. In some cases rating and ranking firms have connections with those they rate or indices they create and license, e.g. MSCI/KLD.

I would add to these six:

- *Delays in reporting*. Some data are available only annually, and there can be delays with obtaining this (dated) data.

- *Lack of auditing of self-reported data*. CSR/sustainability reports are neither standardized nor audited (with a few exceptions).[15]

Another way to put the problem is that on the one hand, there is a range of theorization of what ESG (or CSR/CSP) is, and on the other hand, huge variations in how to measure (and weigh) those factor definitions. These are sometimes referred to as the theorization and the commensurability problems, respectively. Not surprisingly, multiple theorizations and difficult commensurability result in very low 'convergence validity' among and between numerous possible ratings and indexes.[16] The theorization problem is one of *a priori* definitions.

But even with similar definitions, how factors are measured (commensurability) varies widely. Add to this the problem of data (what is being measured, the input), and it becomes clear that there are huge problems.

It is worth noting that few, if any, studies pay detailed attention to what data sources are used, and how. This is a significant gap in the existing literature. This means we simply do not and cannot know what the input information is. Part of this lack of transparency is due to most raters and rankers' protection of their intellectual property and/or protection of data sources. The latter includes proprietary questionnaires sent to firms, and in some cases interviews and/or conversation with firm officials. Thus, the scope of the ESG ratings evaluation problems involves five areas: theorization, commensurability, data, data gathering methods, and transparency.

In turn this directly raises the issue of standardization or lack thereof. Both historically and logically standardization can (and has) come about either through mandates (governmental or otherwise, e.g. stock exchange listing requirements), mergers and market concentration (a few rating firms come to dominate market share), or the creation of public goods standards.

In the latter case, the Sustainability Accounting Standard Board (SASB) comes to mind. Specifically, SASB's focus on materiality based on U.S. legal standards as it defines it from an investor point of view. This clearly does not encompass all stakeholders and stands in

contrast to GRI—the Global Reporting Initiative, for example. Nor does SASB's materiality definition neatly fit into many non-U.S. jurisdictions and regions. SASB's increasing influence in the U.S. and elsewhere does not, however, resolve the data problem: What are the inputs? What SASB's approach does do is solve or radically minimize the commensurability and theorization problems and some elements of transparency due to its clear and granular definition of materiality. We return to this below.

A Deeper Dive: Correlations and information

Correlations

A number of studies have probed the lack of correlations problem, with some of the best focusing on a specific topic.

For example, Semenova and Hassel examined the validity of environmental performance metrics, comparing MSCI (KLD), Asset4 (Thomson Reuters) and Global Engagement Services (GES). While the ratings have common high-level dimensions (theorizations), the authors concluded they do not converge. They argue that the three raters correlate strongly on performance and risk constructs (definitions/theorizations) when they analyzed data for the US MSCI World Index (2003-11), but that there is significant divergence when the authors' analysis controls for company-specific characteristics, e.g. size, profitability.

They conclude that MSCI/KLD 'strengths' focus on company a specific metric while 'concerns' focus on industry wide elements (e.g. emissions, waste, chemical, compliance). Thus, MSI/ KLD data capture historical performance, while GES and Asset4 focus more on environmental opportunity perspectives and future performance metrics.

Semenova's and Hassel's detailed examination concludes that the reason for non-convergence lies in very specific theorizations (compared to high-level agreements), which leads to commensurability issues.[17] They write:

> "We propose that industry-specific environmental risk drives EP [environmental performance] and that performance and risk are different constructs to be clearly separated when conducting further research."[18]

Thus, at a granular level (theorization) there is *a priori* divergence in spite of higher-level agreement. Specifically, they attribute the divergence among the three raters to conflating of environmental impact with environmental management. They also are critical of the lack of control for different rater weightings.[19]

In a similar vein, Dortleitner, Halbritter, and Nguyen examine the overall ESG and the individual E and S and G scores of Bloomberg, Thomson Reuters' Asset4 and KLD (MSCI). Like others, they point out the definitional-theorization problems of CSR/ESG but they additionally focus on selection bias. It's a particularly important factor since, for example, Asset4 relies on firms' disclosure but as well that larger firms with good CSP tend to disclose more than both smaller firms (regardless of CSP) and all firms with poor CSR/ESG performance (however measured).[20] The definition/theorization problem is again typically very granular.

While noting that all three of their raters define E issues along similar lines (e.g. emissions, water, waste, resource reduction, impact of products and services), the authors point out that Asset4 alone singles out animal testing, whereas Bloomberg and KLD (MSCI) focus only on regulatory compliance.

Focusing on S, they write: "... employment quality, occupational health and safety, diversity and opportunity, human rights and product responsibility are primarily evaluated. But when considered in detail, one can see enormous differences in what and to which extent social performance is measured. While KLD evaluates health and safety through only two indicators with regard whether the firm has strong health and safety programs or whether it is involved in controversies, ASSET4 and Bloomberg assess it in a more detailed way."[21]

Their conclusion is worth quoting in detail:

> "The results suggest an evident lack in the convergence of ESG measurements.
>
> First, the qualitative evaluation of the three rating methodologies reveals obvious distinctions in the scoring approaches as well as the CSR definition. This does not only lead to differences in the complexity of the CSP assessment but also in the degree of transparency. While ASSET4 sheds light on various issues regarding social responsibility through qualitative and quantitative questions, KLD combines multiple aspects in one indicator without reporting upon the specific assessment criteria. Although the CSP concepts of the three providers are generally based on similar aspects regarding environmental, social and governance dimensions, the different composition and weighting of the indicators lead to significant distinctions in the final ESG appraisal. This is especially true for the social pillar.
>
> Second, the descriptive statistics confirm the obvious differences in level between the three ESG score providers. Owing to enhanced reporting activities, large firms generally obtain higher scores. By performing a quartile analysis, we find that ASSET4 and Bloomberg ratings are, to a large extent, in line for approximately half of the companies. KLD only shares the same quartile groups for about one-third of the firms. The correlation analysis provides evidence of the fact that the total ESG scores of ASSET4, Bloomberg and KLD are significantly positively correlated with regard to their environmental and social scores. However, KLD shows low correlations to the scores of the other providers. In terms of the particular ESG

dimensions, corporate governance aspects are least strongly
connected to the other pillars.

Third and lastly, the ESG risk analysis demonstrates that the
expected loss is highly dependent on the underlying data basis.
Thus, hardly any correlation exists between the different data
sets in terms of ESG risk. Similarly to the ESG level analysis, the
ESG risk of KLD's ESG scores show the least resemblance to the
other databases. Overall, the results do not indicate a remarkable
coincidence between the expected losses of
the three data providers."[22]

In perhaps the most methodologically-developed study, Chatterji, Durand, Levine and
Touboul examine three raters (Asset4, Innovest and KLD (MSCI) and three indices (Dow Jones
Sustainable Index, FTSE4Good and Calvert). It should be noted that while it wasn't
considered in the latter group, MSCI/KLD is also an index provider. Their data covers
2002-2010 (prior to the KLD purchase by MSCI).

They, too, begin by asking 'how much do we really know about… CSR?' Their conclusion: we
don't really know very much, because the convergent validity among these six is low. It is low
not only due to different theorization, but because 'all or almost all of the ratings have low
validity,' even once one adjusts for differences in theorization.

In other words, commensurability (measurement) is low as well. For example, Chatterji et al.
cite in the E area that KLD gives credit for products beneficial to the environment, while
FTSE4Good uses metrics assessing procedures to find and mitigate environmental risks.
These become difficult if not impossible to compare. Among the six raters in their analysis,
the overlaps run from 19% (Calvert) to 60% (Innovest). An overlap is where all six firms include
the same category. But this itself is misleading as the universes for each of the six vary in
size (as well as composition in other regards, e.g. capitalization). This causes methodological
problems, as I will note below.

"In addition to theorization, commensurability and universe problems,
each rater has its own idiosyncratic weightings. For example, KLD had
71% of its sub-categories in the social area, giving social a higher
preponderance than did Asset4 that had 47% in the social issues
area. The reverse is true when considering employee/human capital
issues; Asset4 puts more weight on it than did KLD."[23]

How does a researcher make sense of and analyze what is known as dichotomous data, a
problem that runs throughout rater and ratings comparisons?[24] Without going into the
methodological particulars, various techniques can begin to sort out these problems,
beginning with joint probability of agreement, Pearson and Spearman correlations and

concluding with pairwise tetrachoric[25] correlations among the six indexes.

Chatterji et al's analysis found pairwise tetrachoric correlations for three years among the six raters, with a mean correlation of 0.30 (about 2 standard deviations). However, this also included some negative ones' correlations, meaning what one rater found responsible another found 'irresponsible.'[26] Correlations (all types) were higher among U.S. based raters compared with European ones, a geographic discrepancy found by many researchers. Most agree this stems from definitional practice/theorization and related weighting differences, as noted in the example above, where KLD (U.S.) weighted social higher than Asset4 (European), which weighted employment higher.

Chatterji et al's final exercise was, after normalizing theorization differences as best they could, to measure commensurability. They found it to be low: That is, measurement and methods themselves were different.[27] While an important finding in itself, what is not examined in their study (and probably not knowable at all) is what data is being measured, that is, what the data inputs are among raters. That is, far more often than not, we simply don't know the inputs being measured or evaluated. This brings us to the information problem.

Information

The organization SustainAbility (http://sustainability.com) conducted a five-part study beginning in 2010 and ending in 2012 called 'Rate the Raters,' the motivation of which was to cast a critical eye on the state of the ESG rating world. Many of the concerns and criticisms they highlighted had previously been noted by others, and would be significantly developed subsequent to their reports.

They summarized their findings, writing that long-standing problems were, "...poor transparency in the ratings process, inadequate focus on material issues, difficulty in comparing companies across industries, [and] conflicts of interest in organizations that offer services (alongside ratings) ..." Additionally, there was 'too much noise' amidst the signals; '...simply too many rating schemes measuring similar things in different ways.'[28]

The reports pointed out a number of related concerns focused on information. There was no standard among raters for distinguishing disclosure (e.g. of carbon) vs. performance (e.g. trends to lower carbon emissions), and no consistency as to how (or whether) to weigh each.

They found about 60% of raters in 2010 depended overwhelmingly on corporate self-disclosure, either in CSR reports or similar publications, or in response to requests for information and interviews from firms. Indeed, they also found that firms that responded to information requests fared better in ratings than firms that did not respond (selection bias). Of the 120 raters they examined, about half relied only on public information sources, and the other half on either corporate self-disclosures alone or some combination of the two.[29] They

found only 'a few' raters adequately disclosed information (sources and methods) so that users could understand how the ratings were constructed.[30]

SustainAbility's critique of the information problem logically led them to conclude that the rating industry must get beyond the 'black boxes' of information. They recognize the intellectual property of these raters resides in their proprietary data sources and methods, but that is of little solace to the users and consumers of such ratings, as they have tremendous difficulty trying to interpret raters' typically differing conclusions and recommendations.[31]

Conclusion

This brief overview of the problem of raters and ratings being all over the map suggests a number of directions for improvement, if not transformation. The four most obvious and most mentioned are: 1-transparency and disclosure regarding both theorization (often, but not always, disclosed at various levels of granularity) and; 2-commensurability (far less disclosed in adequate detail). And 3-less discussed in the literature is data inputs. Finally, 4-how data is interpreted and weighed. As noted above, to varying degrees, each runs directly into the problem of most rating firms' intellectual property (IP). There is no easy way around the IP protection problem given most of the existing rating and ranking models.

Additionally, as noted, there is no existing standard for theorization. On the one hand this makes sense as there are a variety of stakeholders (and shareowners) whose values, orientations, and interests are often quite different from one another.

However, from an investor perspective generally, there are signs of significant change, in particular the growing influence of SASB, both in the U.S. and as its influence grows in other countries and regions. If SASB (or others who may come along) can develop a widely accepted (or mandated) set of specific materiality-based standards as a public good, the focus will subsequently be on data inputs and analyses into these financially oriented materiality categories for what appears to be a more unified financial-investor client audience.

From a broader point of view, however, what is materially important to investors changes over time, and for very good reasons may not be identical to what is important to various stakeholders aside from investors.[32]

A corollary to the standardization argument is the likelihood that the terrain of ratings and ranking competition will shift form theorization to commensurability. If this is correct, the field is open to a paradigm shift in how ESG is conceived, analyzed and measured.

This does not imply that now or in the future there will be a single 'ESG truth', just as there is not and never will be one way to analyze the 'hard' financials of an investment, nor to conceptualize and measure risk and return. That is part of what makes a market work: that players have different views of past and future, and of what is important. What, however, is likely to happen, is that the transparency of concepts and the knowledge of data will dramatically and perhaps qualitatively transform.

In addition to the standardization trend, various forms of big data analytics and artificial intelligence will also drive ESG transformation as these techniques move from a host of non-ESG spaces into ESG and related space. Standardization and technological disruption will

likely transform the rating and ranking industry, resulting in closer (but far from perfect) convergence. The hallmark indicator of this transformation, should it fully take place, will be transparency in the theorization and commensurability realms, and importantly in the transparency of data inputs, which, as noted, is currently lacking in traditional rating and ranking organizations.

Moreover, while all rating and ranking organizations depend on existing data (regardless of how they access and analyze it), in the near future it is likely that firms will begin to create data using advancements in and declines of cost for various technological advances able to supply data which currently does not exist.

For example, satellite data is currently tracking various climate indicators, but it could also track parts of the global supply chain, creating a new S source. Similarly, advances in so-called asset specific data can layer multiple of data sources to identify a specific firm's (or its sub-contractors) behavior regarding both S and E issues. A number of academic conferences have been held or will be examining these technological potentials.[33]

In sum, there will always be important and legitimate differences of evaluating, measuring and defining E and S and G. But the investors or data users, not the raters/information providers, should determine those differences.

The data and methods of raters need to be transparent and be able to be manipulated by the end user to fit his or her quantitative and/or qualitative models. This means that data should be available in 'raw' form, much like financial data (and the SASB model). The data inputs may not be 'objective' (that is, how an event is covered and/or interpreted by a news story, an NGO, a corporate, an investor, etc.), but it must be transparent.[34] It likely goes without saying, or should, that technological advances have the potential to make analysis and data available in near real time unlike most current raters and rankers. This opens up the ability to use a variety of statistical tools that are not meaningful when reports are mostly issued annually.

Finally, it is worth highlighting that the Global Initiative for Sustainability Ratings project's 'core framework' emphasizes many similar themes. Best they speak for themselves:[35]

GISR'S 12 Principles of the CORE Framework

C◆RE	Framework
Process	
Transparency	A rating should be transparent to those whose decisions are affected by the application of such rating.
Impartiality	The design and application of a rating, whose primary users are external to the evaluated company, should be protected from undue influence by such company.
Continuous Improvement	Through periodic update, a rating should track and integrate the best-available science, measurement techniques, issues and indicators.
Inclusiveness	Development of a rating should identify and systematically engage those stakeholders whose decisions are influenced by the application of the rating.
Assurability	A rating should be designed to allow for independent, third-party assurance that its application comports with the GISR Principles.
Content	
Materiality	A rating should assess performance based on sustainability issues relevant to the decision-making of stakeholders for which a rating is designed.
Comprehensiveness	Evaluating one or more aspects of sustainability performance should systematically assess for impacts on human, intellectual, natural and social capital.
Sustainability Context	A rating should assess performance in the context of science-based thresholds and limits, or, if unavailable, widely-accepted norms pertaining to long-term human and ecological well-being.
Long-Term Horizon	A rating should enable the evaluation of the long-term performance of a company while simultaneously providing insights into short- and medium-term outcomes in alignment with the long-term.
Value Chain	A rating should reflect all portions of a company's value chain over which the company exercises significant influence.
Balance	A rating should utilize a mix of measurement techniques to capture historical and prospective performance.
Comparability	A rating should allow users to compare the performance of the same company over time and of different companies within the same time period.

End Notes

[1] Russ Kerber and Michael Flaherty, "Investing with green ratings? A gray area," *Reuters*, accessed July 2017, https://www.reuters.com/article/us-climate-ratings-analysis-idUSKBN19H0DM.

[2] Iris H-Y Chiu, "Standardization in corporate social responsibility reporting and universalist concepts of CSR?" *Florida Journal of International Law*, vol. 22 (2010).

[3] Asset4 has been part of Thomson-Reuters since 2009.

[4] KLD (Kinder, Lydenberg, Domini), the original index of Domini Social Investments, a 400-firm cap-weighted index focusing on E and S (and at first less on G) issues. The most widely used index/benchmark, e.g. by BlackRock, MSCI. (https://www.msci.com/documents/10199/904492e6-527e-4d64-9904-c710bf1533c6). Like many older indices its metrics, weightings and categories have changed over time.

[5] Timothy A. Hart and Mark Sharfman, "Assessing the concurrent validity of the revised Kinder, Lydenburg and Domini corporate social performance indicators", *Business and Society*, vol. 54 (2015): 583.

[6] 'Who are the ESG rating agencies' (*Sustainable Perspective*, February 2016). Accessed July 2017, https://www.sicm.com/docs/who-rates.pdf.

[7] Miles Livingston, Jie (Dianna) Wei, Lei Zhou, "Moody's and S&P Ratings: Are They Equivalent? Conservative Ratings and Split Rated Bond Yields," *Journal of Money, Credit and Banking* (42:7), Oct. 2010, 1267-1293.

[8] See, Richard Cantor and Frank Packer, "Difference of opinion and selection bias in the credit rating industry," *Journal of Banking and Finance*, 21:10 (October 1997), 1395-1417; and, Yoon S. Shin and William T. Moore, "Explaining credit rating differences between Japanese and U.S. agencies," *Review of Financial Economics*, 12:4 (2003), 327-344.

[9] Global Initiative for Sustainability Ratings. Accessed July 2017, http://ratesustainability.org/hub/index.php/search/report-in-graph. From these hundreds of organizations there are about 10,000 different ESG/sustainability KPIs (key performance indicators). Others estimate that there were in 2015 about 500 rankings, 170 ESG indices, over 100 ESG type awards, and at least 120 voluntary ESG/sustainability standards. (Stephanie Mooji. "The ESG Rating and ranking industry." Working paper, Oxford University, Smith School of Enterprise and the Environment, April 11, 2017.)

[10] Remco van den Heuvel. "How Robust are CSR Benchmarks?" (Master's thesis, Tilburg University, 2012.)

[11] The E concern focused on clear cutting rainforests with significant impact on endangered species, among other effects. The S concern focuses on what some (mostly E.U. authorities) see as potentially significant when the oil is not processed effectively.

[12] Eliot Caroom, "ESG and Rater Subjectivity." *Lipper Alpha blog. Accessed* July 2017, http://lipperalpha.financial.thomsonreuters.com/2016/11/failing-scores-for-esg-raters/

[13] Thomas Seubert, "Are standardized ESG disclosures effective?," *WealthManagement.com.* Accessed July 2017, http://www.wealthmanagement.com/mutual-funds/are-standardized-esg-disclosures-effective

[14] Sarah Elena Windolph, "Assessing corporate sustainability though ratings: Challenges and their causes," *Journal of Environmental Sustainability*. 1:1 (2011).

15 Olivier Boiral, "Sustainability reports as simulacra? A counter-account of A and A- GRI reports," *Accounting, Auditing and Accountability*. 26:7 (2013), 1036-71. which analyzes what the authors sees as deeply flawed high ratings, based on self-reports, from GRI (Global Reporting Initiative).

16 Aaron Chatterji, Rodolphe Durand, David Levine, Samuel Touboul. "Do Ratings of firm converge? Implications for managers, investors and strategy researchers." Social Science Research Network (SSRN.com) id: 2524861. October 2014. Accessed July 2017, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2524861.

17 Natalia Semenova and Lars G. Hassel, "On the validity of environmental performance metrics," *Journal of Business Ethics,* 132: (249-258): 2015.

18 Ibid., 250.

19 Ibid, 252.

20 Gregor Dorfleitner, Gerhard Halbritter, and Mai Nguyen, "Measuring the level and risk of corporate responsibility: An empirical comparison of different ESG rating approaches," *Journal of Asset Management*, vol 16 (450-466): 2015. p.453: Selection bias is widely commented on by analysts.

21 Ibid, 456.

22 Ibid, 465.

23 Aaron Chatterji, op. cit. p.9.

24 Examples dichotomous data among raters and rankers include differing number of firms in each rater's universe, rater's views of cut off levels (to include a firm in its universe. These make apples to apples comparisons difficult or impossible.

25 Tetrachoric correlations are able to estimate '...the quantitative magnitude of the relationship between two raters [that are dichotomous] ... [so that it is] invariant to the number of companies selected in each index...', unlike the more familiar Pearson correlation. (Ibid., 13). For a full explanation see this example: http://www.statisticshowto.com/tetrachoric-correlation/ .

26 Aaron Chatterji et al note there is no clear determination of what is a high or low tetrachoric correlation. Assuming normal data distribution, they suggest that 0.68 is quite strong; 0.45 substantial and 0.40 relatively low, somewhat parallel to how Pearson correlations are seen. (Ibid., 14)

27 Ibid., 20-27.

28 SustainAbility, "Rate the Raters: Phase One," p.2. Accessed July 2017, http://sustainability.com/our-work/reports/rate-the-raters-phase-one/

29 SustainAbility, "Rate the Raters: Phase Two," 5, 9-10. Accessed July 2017, http://sustainability.com/our-work/reports/rate-the-raters-phase-two/

30 Ibid., 10.

31 Ibid., 3.

32 Robert G. Eccles and Tim Youmans, "Materiality in corporate governance: the statement of significant audiences and materiality," *Journal of Applied Corporate Finance*, 28:2 (Spring 2016): 39-45. Jim Hawley, "Is 'materiality' in the eye of the beholder?" Parts 1 and 2, Accessed July 2017, https://blog.insight360.io/is-materiality-in-the-eye-of-the-beholder-part-i-fae14b0f7c24 and https://blog.insight360.io/is-materiality-in-the-eye-of-the-beholder-part-ii-e14e9c72a0e4 .

33 E.g. Yale Center for Business and the Environment forthcoming symposium in September 2017 on "The state of environmental, social and governance (ESG) data and metrics," and the University of Oxford's Smith School of Enterprise and the Environment conference in April 2017 called "From disclosure to data-toward a new consensus for the future of measuring environmental risk and opportunity."

34 In this regard, the recent ratings of mutual funds for their 'ESG' performance is not a particularly helpful trend as it uses, for example, Sustainalytics ratings of Morningstar approximately 20,000 listed mutual funds and ETFs. Similarly, MSCI's ESG scores are also used for rating about 21,000 mutual funds and ETF's. Both implicitly are taken as a 'truth' that is not easily tested. These ratings are at a doubly high level undercutting their meaning: they aggregate not only dozens or hundreds of firms in each fund, but do so only on high-level scores for each firm. (See for contrary views, https://www.msci.com/documents/10199/84bcc5fa-783e-4358-9696-901b5a53db3b; and, Jon Hale, "Does using Sustainalytics data affect the Morningstar sustainability ratings of MSCI-based funds?" Accessed July 2017, https://medium.com/the-esg-advisor/does-using-sustainalytics-data-affect-the-morningstar-sustainability-ratings-of-msci-based-funds-a300973a6073.

35 "About," Accessed July 2017, http://ratesustainability.org/about/