Title

Do Ratings of Firms Converge? Implications for Strategy Research

Authors
Chatterji, Aaron
Durand, Rodolphe
Levine, David
et al.

Publication Date
2014-04-01

IRLE WORKING PAPER
#107-14
April 2014

Do Ratings of Firms Converge? Implications for Strategy Research

Aaron Chatterji, Rodolphe Durand, David Levine, and Samuel Touboul

Cite as: Aaron Chatterji, Rodolphe Durand, David Levine, et al. (2014). “Do Ratings of Firms Converge? Implications for Strategy Research”. IRLE Working Paper No. 107-14.
http://irle.berkeley.edu/workingpapers/107-14.pdf

Do Ratings of Firms Converge? Implications for Strategy Research

Aaron Chatterji
Duke University

Rodolphe Durand
HEC Paris

David Levine
UC-Berkeley

Samuel Touboul
HEC Paris

April 2014

Abstract

Raters of corporations play an important role in assessing domains ranging from sustainability to corporate governance to best workplaces. Scholars increasingly rely on these ratings to test theories about corporate social responsibility (CSR), corporate governance and the influence of stakeholders. Though these raters frequently develop sophisticated methodologies, we find they often diverge in their ratings of the same firm, creating uncertainty for managers and stakeholders, and also posing challenges for researchers. We document the surprising lack of convergence of social ratings for the first time using six well-established socially responsible investing (SRI) raters, with comparisons of overlap, correlations, and regression analysis. Our results suggest that scholars should interpret empirical results with caution and at least use multiple ratings schemes in studies of CSR and governance.

In 2010, professional fund managers in the U.S. invested more than $3 trillion under the banner of socially responsible investing (SRI).[1] The enormous amount of capital allocated to SRI has drawn considerable attention from scholars, activists, managers, and policymakers interested in the drivers of corporate social responsibility (CSR). Some CSR advocates praise SRI, believing that it can direct capital toward the most responsible firms while penalizing firms with poor social performance. Skeptics argue that the organizations that rate the social performance of enterprises, referred to as "raters" or "SRI raters" in our study, cannot discern which firms are socially responsible.[2] For example, Hawken (2004) points out that the various methodologies employed by socially responsible raters allow for almost any public firm to be a member of at least one SRI index. Entine (2003) presents several examples of raters giving high marks to firms that were later embroiled in famous scandals. Delmas, Etzion and Nair Birch (2013) show that different raters may use different methods to measure firms' environmental performance.

Academics have produced dozens of articles on CSR and SRI over the past two decades, with growing interest in recent years (Orlitzky, Schmidt, & Rynes, 2003). For example, from 1994-2008, seven articles published in SMJ relied on KLD data. From 2009 to 2013, 19 articles used KLD and 6 articles employed FTSE4Good, Innovest, DJSI or Asset4. Notably, influential research has examined the effects of SRI on returns for investors and the cost of capital for managers (Galema, Plantinga, & Scholtens, 2008; Waddock, 2003). Other research examines the drivers of CSR, such as profit-maximizing responses to heterogeneous consumer preferences (Mackey, Mackey, & Barney, 2007), imitation among firms, or a departure from profit-maximizing behavior to satisfy managers' private goals (Marquis, Glynn, & Davis, 2007; Devinney, 2009).

[1] Social Investment Forum Foundation, Report on Socially Responsible Investing Trends in the United States, 2010. According to this source, as of 2010, socially responsible investments are nearly 12.2% of the total funds managed by professional investors. This percentage has grown markedly since 2005, where $2 trillion, or 10% of total funds, were invested in accordance with socially responsible guidelines.

[2] We use the term "raters" or "SRI raters" to refer organizations that assess corporate social responsibility.

A key question of this study is whether raters converge to valid assessments of firms' social activities and performance.[3] Despite growing interest in CSR, little research examines whether raters measure CSR accurately (Sharfman,1996; Delmas et al, 2013). If these metrics are invalid or are inconsistently applied across raters, scholars who conduct analysis using one rating scheme risk drawing conclusions that are not generalizable. Lack of convergence among raters would also pose significant challenges for practice. Socially responsible firms seeking to improve their CSR should be able to understand whether poor ratings are due to poor results, a different conceptualization of CSR than the raters, or poor measurement methods (Margolis & Walsh, 2003; Gray, 2010). Furthermore, if ratings cannot consistently identify socially responsible firms, the hypothesized benefits of SRI cannot occur. In the worst-case scenario, if firms expend resources to achieve high scores on invalid metrics, then even well-intended attention to social metrics reduces social welfare. Thus, it is crucial, both from the academic and practical perspective, to understand the validity of social ratings and the dynamics driving convergence across raters.

In this paper, we first document that the ratings of six major social raters—KLD, Asset4, Calvert, FTSE4Good, DJSI, and Innovest—have little overlap in membership and fairly low correlations with each other. Our results imply that SRI raters not only do not agree on a one definition of sustainability (their "theorization" of CSR or what they measure), but also that raters may measure the same construct in different ways (the "commensurability" of CSR dimensions or how raters measure the same indicators). The validity of social ratings is a serious concern not just for academics, but also for investors, activists, and policymakers. Our findings suggest scholars should interpret prior empirical studies using CSR ratings with appropriate caution and at the very least, replicate studies using alternative ratings schemes.

[3] When discussing the behavior of raters, we use the term "convergence." When referring to the rating they provide, we use the term "convergent validity."

APPROACHING CONVERGENCE

While we have broad agreement in the field on how to measure financial performance, assessing social performance is inherently more challenging. The literature on social evaluations of firms and organizations establishes that two mechanisms drive convergence. First, "theorization" makes clear precisely what raters assess and why it matters (Durand, Rao, & Monin, 2007; Hsu, Roberts, & Swaminathan, 2012). Next, "commensurability" of indicators makes comparison across evaluated organizations possible (Espeland & Sauder, 2007; Sauder & Espeland, 2009).

"Theorization", according to Rao et al. (2003), is the conceptual discourse produced by a rater (e.g. Michelin in haute cuisine, US News in higher education) that associates actions to outcomes and allows organizations to expect (1) better rankings from changes in behavior and (2) the accompanying benefits from these changes, such as more customers. When there is a clear theorization, organizations can adjust their behaviors—or choose not to. We use the term "theorization" to refer to the beliefs raters may have about what being socially responsible means. A "common theorization" refers to agreement across raters on a common definition of CSR; for example about dimensions of social investors should care about (e.g. environmental, social, and corporate governance), or about industries that social investors should consider as inherently irresponsible (e.g. nuclear energy, weapons, tobacco).

"Commensurability" of a construct is high when different raters measure the same construct in a similar fashion. For instance, in financial ratings, measurements and interpretation of the construct "debt/equity ratio" are similar across various rating agencies. We use the term "commensurability" to refer to the extent that raters are using the same (or at least similar) measures and methods to assess the same construct (e.g. employee safety or independent board).

Simply put, common theorization among SRI raters is overlap in what raters choose to measure, and commensurability is overlap in how they measure corporate social responsibility. In any given domain, raters are more likely to converge around valid measures when the raters share a same theory of what good performance means ("theorization") and what indicators are valid proxies for that good performance ("commensurability").

Common theorization

When evaluating the extent of common theorization across SRI raters, there are at least three aspects of measurement to consider. First, what high-level categories (e.g., environmental, social, governance) do the raters measure? Second, do the raters screen out particular industries such as tobacco and firearms? Third, do raters normalize their ratings by industry such that a firm is compared to the other firms in its own industry?

In terms of high-level categories, there is broad agreement on the components of social responsibility. Rhetorically, the marketing materials of the raters we study all seem fairly similar in describing their goals. For example, one of FTSE4Good's stated goals is "to provide investors with the opportunity to gain exposure to companies that meet globally recognized corporate responsibility standards."[4] KLD asserts that its "research is designed for investors and money managers who integrate environmental, social and governance factors into their investment process."[4] Calvert describes its ratings as "a broad-based, rigorously constructed benchmark for measuring the performance of large, US based companies following sustainable and responsible policies…"[4], and Asset4 claims to "provide objective, relevant and systematic environmental, social and governance information" that "professional investors use to define a wide range of responsible investment strategies."[4] In addition, all of the indexes cover similar high-level topics, including environmental and social performance.

[4] While our empirical analysis utilizes data from 2002-2010, we have tried to provide more recent information where possible, including: FTSE4Good Index Series http://www.ftse.com/Indices/FTSE4Good_Index_Series/_Downloads/Brochure__english.pdf (Last accessed March 1st, 2012); KLD's Research Products http://www.kld.com/research/index.html (Last accessed August 13th, 2007); Calvert-About the Ratings http://www.calvert.com/sri-index.html (Last accessed March 1st, 2012); Asset4 ESG content overview http://thomsonreuters.com/products_services/financial/content_news/_content_overview/content_az/content_esg/ (Last accessed February 8th, 2012).

However, there are some key differences across the raters. Some raters consider additional high-level categories. For example, KLD and Asset4 rate firms according to their products' safety, while other raters do not. Asset4 and DJSI explicitly consider economic dimensions, while other raters do not. KLD, Asset4, FTSE4Good and Innovest consider Corporate Governance as part of CSR while Calvert and DJSI do not. Interestingly, the geographic origin of the rater appears to have some influence on their theorization of CSR. As an example, KLD, a U.S. rater, has 71% of its sub-categories[5] in the social issues domain. KLD therefore puts more weight on social issues than Asset4, a European rater, which has only 47% of its sub- categories[6] related to social issues. In other domains, such as in issues relating to employees, Asset4 appears to place more emphasis as compared to KLD. While both Asset4 and KLD consider employee diversity, the firm's impact on local communities and its respect of human rights, Asset4 clearly differentiates between employees' health and safety, training programs, and labor relations. KLD includes all of those topics under the broad umbrella of "employment".

Further differences in theorization appear when considering the use of screens for particular industries. Three of the six raters (KLD, Calvert, and FTSE4Good) use explicit screens to exclude firms with substantial investments in categories like tobacco and firearms, though they each define "substantial" differently. Even among this group, FTSE4Good and KLD screen out firms involved in nuclear power, while Calvert does not. Finally, four of the six raters normalize their ratings by industries (KLD and Asset4 are the exceptions). These raters assert that CSR performance must be measured relative to industry peers (see Table 1)

Insert Table 1 about here

[5] Community, Governance, Diversity, Employment, Environment, Human Rights, Product.

[6] Function of the board of directors, Structure of the board of directors, Compensation of the board of directors, Vision and strategy, Shareholders, Emission reduction, Product Innovation, Resource Reduction, Product Responsibility, Community, Human Rights, Diversity, Employment Quality, Health and safety, Training and development

The upshot is that despite similar language there do appear to be differences in the way various raters envision CSR and which firms should be evaluated in the first place.

Commensurability

Low convergent validity due to lack of common theorization is still consistent with high validity of raters, if each of them is trying to measure a different definition of "good CSR." For example, it is not a critique of either rater if the list of "100 best cheap eats" and "100 best fine dining" do not overlap, as each has a different theory of what diners are looking for. Similarly, users of social ratings may differ in what dimensions of CSR they value (Crilly, Zollo, & Hansen, 2012; Delmas & Toffel, 2008; Philippe & Durand, 2011). Some investors may wish to avoid profiting from activities they feel are harmful, leading them to desire screens based on whether a firm sells certain products. Other investors may wish to encourage high effort by managers, leading them to focus on ratings that are defined relative to an industry, not an absolute scale. In that case, low correlations across social ratings could still be consistent with valid measurement by each rater, because raters would be simply appealing to different groups.

However low convergent validity will still be present in the case of low commensurability across raters, or when ratings of the same construct disagree due to differences in measurement. Thus if we adjust for different theorizations (what constructs raters measure), the convergent validity of ratings will be determined by differences in commensurability (how raters measure the same constructs). Commensurability is inherently a serious challenge for SRI raters. For example, it is unclear exactly how to measure superior human resource management, or which indicators to use to measure higher-than-average toxic releases. Similarly, raters must quantify information that is difficult to measure, such as the social impact of additional minority representation on the board of directors, or the social impact of having business interests in a nation that is ruled by totalitarian regime.

Raters make a significant effort to persuade potential investors that their methods and ratings are based on careful analysis of high-quality data (Chatterji, Levine, & Toffel, 2009). The implication is that they measure the indicated constructs with high validity. For example, all of the social raters claim they draw on multiple sources and use multiple research methods, both of which are established scientific approaches: They all review official government data (e.g., on toxic emissions and regulatory actions), explore company documents and press reports, and conduct interviews. Our research confirms that all the raters (except Asset4) also do surveys, though they employ different methodologies. All of these raters' have marketing materials that stress how carefully they analyze companies' social, governance, and environmental records. They often compare themselves to traditional financial research firms. For example, KLD describes its services as "analogous to those provided by financial research service firms." Not coincidently, Dow Jones and the Financial Times (Creators of DJSI and FTSE4Good) and Thomson-Reuters (owner of Asset4) are also well-known providers of traditional financial information.

Nevertheless, raters use different methods and variables to measure the same construct. Some raters measure environmental performance with indicators of a firm's environmental processes, while others will concentrate on the firm's environmental outcomes (Delmas et al., 2013). For example, raters such as KLD give credit for products with beneficial impact on the environment, while others, like FTSE4Good, employ metrics that assess the procedures to identify and fix environmental hazards, in the spirit of the ISO 14001 management standards. In general, these differences in commensurability are difficult for investors to observe.

In sum, there are two possibilities regarding convergent validity of SRI ratings after adjusting for theorization. If commensurability is high, adjusting for different theorizations should substantially increase convergent validity. For example, if all raters measure environmental performance in the same way, convergent validity should be high. Alternatively, it is possible that the raters may themselves be uncertain about how to accurately measure each dimension of social responsibility.

Hence, we might expect that even after adjusting for differences in theorization, convergent validity will remain low. In this case, if convergent validity is low for a pair of raters rating the same constructs, at least one of the raters has low validity as well. Below, we perform these tests to assess the convergence of SRI raters.

DATA

To test the convergence of SRI raters, we examine the ratings of a common universe of companies from six leading social raters: KLD, Asset4, Innovest, DJSI, FTSE4Good and Calvert. Taken together, these raters and ratings are among the most popular and well established in the field.[7] These data cover the 2002–2010 period for KLD and Asset4. For the other raters we have selected years: 2004 for DJSI, 2005 for Calvert and Innovest, and 2006 for FTSE4Good. In all instances, we compare ratings provided in the same year, unless otherwise noted. Our dataset provides a global view of the industry, with KLD, Calvert, and Dow Jones based in the U.S., Innovest in Canada, while FTSE4Good and Asset4 have origins in the European Union.[8] The raters have broadly similar processes to develop ratings. They collect raw quantitative and qualitative data on specific information (production of tobacco based products, CO^2 emissions, election of trade-union representatives, etc.). The raters then implement proprietary methodologies to issue scores on high-level categories such as environmental impact, human rights compliance, and governance. Finally, raters typically provide a list of companies they consider most responsible, most often in an equity index for potential investors.

To assemble the data, we started with each rater's index of socially responsible companies and the broader universe of company stocks from which the index list was selected (S&P500, Russell 1000). Our first task was to denote the firms that had been included on each rater's index of top

[7] SustainAbility report, Rate the Raters Phase Two, Taking Inventory of the Ratings Universe, 2010. This report lists all of these raters, except for Calvert, among their top 16 raters in terms of credibility. Note that KLD purchased Innovest at the time of this report. We included Calvert since it is regarded as one of the oldest and most well-known raters in this space

[8] FTSE4Good is based in the UK, while Asset4 is in Switzerland.

social investments. Thus, we assigned a "1" to firms included in the KLD Domini 400 Social Index, the Calvert Social Index, the FTSE4Good Index, the DJSI World Index, Innovest's 18 U.S.-based firms in its "Top 100 Leaders in Sustainability," and Asset4 firms which received an A+ grade. We assigned a "0" to firms in the eligible universe but not in these indexes. In sum, we obtained membership data for 3134 firms from six different indexes' universes. The universe common to all raters includes 551 firms in 2004, 413 in 2005 and 538 in 2006, and is most comparable to the S&P 500. Table A1 in the Appendix summarizes the raters' universes.

In addition to membership, we collected more detailed data for all firms rated by KLD and Asset4 between 2002 and 2010, and for some firms rated by Calvert and Innovest in 2005, and by DJSI in 2004. For KLD, we had 98 detailed sub-scores, which rated each company on more specific aspects of their environmental and social performance. The KLD sub-scores consist of 1/0 indicators for a strength or a concern on topics such as waste recycling, involvement in military products, and emissions of ozone-depleting gases. Those strengths and concerns are grouped in 7 categories (Environment, Community, etc.).[9] We used these sub-scores in two different ways. First we computed the sum of strengths minus the sum of concerns per category. Second, we estimated KLD category scores with the predictions from of a logit model that considered membership to KLD DS400 as a binary dependent variable, and KLD strengths and concerns per category as independent variables. We refer to this second measure of KLD scores as "the probability of inclusion in DS400". For Asset4 we accessed scores for the four high-level categories and corresponding 18 sub-scores.[10]

[9] Community, Diversity, Employment, Corporate Governance, Environment, Human Rights, Products.

[10] Economic (Economic Performance, Shareholders' Loyalty, Clients Loyalty), Governance (Board Functions, Board Structure, Compensation Policy, Vision and Strategy, Shareholder Rights), Environment (Emission Reduction, Product Innovation, Resource Reduction), Social (Product Responsibility, Community, Human Rights, Diversity and Opportunity, Employment Quality, Health & Safety, Training and Development)

We had fewer details on other raters' sub-scores. For Calvert, we had five high-level scores[11], but only for the 100 largest firms they rate. For DJSI, we had scores for its three high-level categories and for 78 firms which represented the within-industry top 10% of firms plus one "runner-up" per industry. Innovest computes its index by first issuing each firm a numerical score, which is then normalized per industry to become a letter grade (AAA down to CCC). This letter grade is used as an indication of index membership. We had access to Innovest's letter grades for each firm in their universe and for three high-level categories (Social, Environment, and Governance). We transformed those grades into a 1 to 7 score for our analysis.

METHODS AND RESULTS

We first explore overlap among raters in terms of their assessments of CSR. In the Appendix, Table A2 shows that several well-known firms are included in some raters' social indexes, but not in the others. Google, for example, was only considered as socially responsible by Calvert in 2005. However, does this indicate that Google is not socially responsible? Or alternatively, that its theorization of CSR is close to Calvert's? Or finally that Calvert measures CSR in a way that advantages Google? By analogy, if only one financial analyst included Google in a preferred stock portfolio, would this indicate a poor financial outlook for Google, or just divergent preferences of the financial analyst? Table A2 provides initial insights about the low convergence of SRI raters. Strikingly, in 2004 only six companies[12] are either in all or none of the most popular SRI raters' indexes. To make a more careful assessment, we can also explore convergence by measuring the likelihood that a company included in KLD's DS400 is also included in the DJSI. In doing this exercise, we must take into account that the raters' universes may differ: e.g. KLD only rates firms based in the US, and thus European firms contained in Asset4 index are not eligible for the KLD

[11] Environment, Workplace, Business Practices, Human Rights, and Community Relations

[12] Google, Procter and Gamble, Walmart, UPS, Valero Energy, & Bank of America

index. Taking into account common universes, results from Table 2 provide further insight into the low convergence of SRI raters, with an average overlap between indexes ranging from 19% to 60%.

Insert Table 2 about here

However measuring convergent validity more rigorously across six raters is a challenging exercise. To properly assess the convergent validity of SRI ratings, measuring overlaps is not enough, and several methodological choices have to be made. There are numerous measures of similarity among discrete and continuous ratings. The most common of them are the joint probability of agreement, the kappa statistics, and the Pearson and Spearman correlations. However, none of them are appropriate for our setting. In our case, examining the share of overlapping membership between pairs of indexes can be misleading as each index does not include the same number of firms. For example, if one index includes 500 firms from a universe of 1000 and a second index includes only 10 of that universe, it would be surprising if almost all of the second index's members were not in the first index. Secondly, statistical significance can be a misleading indicator of convergent validity when the null hypothesis is zero relation between the two ratings. Convergent validity requires a stronger relationship than just an association different from zero, and we need measures that not only test the statistical significance of the relationship, but also its magnitude.

We therefore measure the convergent validity of ratings by examining the pairwise tetrachoric correlations between the six indexes. Tetrachoric correlation is a maximum likelihood technique that estimates the correlation of two raters' unobserved continuous ratings on entities when only a discrete membership is observed. This measure is a correlation adjusted for the dichotomous nature of the data and the cutoff level of each rater (see Appendix for further details). As an illustrative example, consider two psychiatrics who analyze the same population. Even if their assessment of patients' degree of depression is identical, they perceive different cutoffs of when drugs are effective, so they do not prescribe drug therapy to the same number of people. This "membership" depends on their cutoff point, below which they believe the patient does not require drug therapy. In such a case the Pearson or Spearman correlations between treated and not treated patients will be low, while the tetrachoric correlation will score high. The intuition behind the cutoff is as follows: Given the assumed normal distribution of the continuous score, the % of "approved" firms implies each rater's cutoff. For example, if 50% of firms are listed as "approved", then the cutoff is the median. If approximately 2.5% of firms are listed as "approved", the cutoff is 2 s.d. greater than the median.

Thus, pairwise tetrachoric correlations provide us with a more precise assessment of the quantitative magnitude of the relationship between two raters, and is invariant to the number of companies selected in each index. Pairwise tetrachoric correlations in 2004, 2005 and 2006 between the six raters on the universe common to each pair of raters are presented in Table 3.

Insert Table 3 about here

Mean correlations between a given index and the other raters' indexes range from 0.13 to 0.52, which indicates low convergent validity among raters. By comparison, substantial agreement is typically ascribed to values above 0.6 for Cohen's Kappa or Fleiss' Kappa (Landis and Koch, 1977). Overall, the tetrachoric correlations between pairs of indexes are fairly low. They range from -0.12 between Calvert and Asset4 A+ in 2005, to 0.67 between Innovest and Asset4 A+ in 2005. Only 3 of the 12 correlations are higher than 0.5. Negative correlations between several indexes indicate disagreement among raters: in such cases, members of the first index have a greater chance of being non-members than members in the second index.

However, while overall convergence is low, some similarities exist between groups of raters, specifically between raters based in the U.S (KLD, DJSI, Calvert) and raters based in the European Union (FTSE4Good, Asset4). The average tetrachoric correlations between US raters (0.45) and between EU raters (0.53) are higher than the average correlation between all raters (0.31), suggesting some limited evidence that geographically proximate raters may have closer theorizations of CSR.

Our key results are robust when other KLD indexes such as KLD BMS or KLD LCS are taken into account (see Appendix A3 and A4, Panel A). Further our results hold when we examine only the sub-group of firms that are common to every rater's universe. Average correlations between indexes range from 0.08 to 0.44; the average tetrachoric correlation of US based raters reaches 0.47, for EU based raters it is 0.54, and their overall average correlation is 0.30 (see Appendix A5).

We also explore the tetrachoric correlations between KLD DS400 and Asset4 A+ over time on the overlapping universe of firms: 0.08 (2003), 0.26 (2004), 0.08 (2005) and 0.14 (2006), showing no evidence that convergent validity is improving (See Appendix A4, Panel B). This pattern is also apparent using data from KLD BMS (see Appendix A5). Taken together, the low tetrachoric correlations between the six SRI raters, and the lack of improvement over time between KLD DS400 and Asset4 A+ provides further evidence that there is low overall convergent validity among SRI ratings.

Adjusting for Differences in Theorization

Next, we adjust for explicit differences in theorization among raters. Our adjustment builds on Asset4's continuous "social responsibility" score for each company it rates. If Asset4 and another rater have similar theorization and high commensurability, then members in the other rater's socially responsible index will have much higher Asset4 scores than non-members. At the same time, it is possible that some highly rated Asset4 firms are not in the other rater's index because the other rater uses a screen (e.g., tobacco) not used by Asset4 (which uses no screens). In that case members of the other rater's index may not have a higher Asset4 scores than non-members. However, we can adjust for screening and normalizing procedures and explore again whether members in the other rater's index have higher Asset4 scores than non-members.

Our methodology follows this rationale. We first standardize Asset4 continuous scores ($R_{iAsset4}$) so that they have a zero mean and a standard deviation of one. We then compute the difference in the means of Asset4 continuous scores between members and non-members of each of the six indexes. Those "membership gaps" are computed for each index *i* as follow:

$$Membership\ Gap_i = \frac{\sum_{c\ in\ index\ i} S_c}{m} - \frac{\sum_{c\ not\ in\ index\ i} S_c}{n-m}$$ where:

- *c* indexes companies in the universe *n* shared by rater *i* and Asset4
- *m* is the number of firms in the index of rater *i* within *n*, the overlapping universe
- S_c is the standardization of R_c, the Asset4 score for company *c*.

Insert Table 4 about here

The top row of the top panel of Table 4 shows the membership gaps. They measure whether membership in one of the SRI indexes is a good predictor of the Asset4 continuous score. If raters were to have a same theorization and commensurability (the same Ri as Asset4), these gaps should discriminate equally and hence have similar values. However, while the gap between Asset4 Index members and non-members equals 1.80 standard deviations in 2006, for this same year, the gap between members and non-members of the FTSE4Good index is only 0.90 standard deviations, and 0.26 for KLD-DS400. Members of the Calvert index even have Asset4 continuous scores significantly below the non-members (with a gap of -0.21 standard deviations compared to the Asset4 gap of 1.82 in 2005), providing evidence of no convergent validity between Calvert and Asset4.

Next, we adjust these gaps for differences in industry normalizing and screening.[13] Therefore while the upper row of Table 4 represents the gap in Asset4 scores between members and non-members of each index when differences in theorization are not controlled for, the four lower rows present results when these differences are accounted for. In most cases, the gap between members

[13] For Innovest, DJSI, Calvert, and FTSE4Good styles we mimicked industry normalization by standardizing Asset4 continuous scores per industry, using the first four digits of firms' Thomson Reuters Business Classification code. For KLD, Calvert, and FTSE styles we mimicked screening methodologies by assigning a zero score to firms (before standardization of scores) that did not comply with the specific screening criteria.

and non-members increases and get closer to the recalculated gap for Asset 4. For example, in 2004 the KLD DS400 gap goes from 0.29 to 0.68 when adjusted for KLD's methodology. In doing so, it does get closer to the Asset4 / KLD style gap of 1.31 but still remains quite distant. Although these results provide evidence that different theorizations are partly responsible for the low convergent validity between raters, this convergent validity remains low even after adjusting for explicit differences in theorization. The implication is that low convergent validity between SRI raters is not only driven by different theorizations, but also by low commensurability among most pairs of raters. As a robustness check, we used the same approach with our two measures of KLD continuous scores to assess the convergent validity of other indexes with the KLD DS400 index. We continue to find low convergence among raters, even when adjusting for differences in theorization (See Appendix A6 and Appendix A6bis that uses our two different approaches to KLD scores).

The third condition that explains divergences in rating is based on the non-overlapping aspects of social responsibility that raters choose to measure. For example, all raters consider firms' environmental responsibility, but only Innovest, FTSE4Good, Asset4, and KLD evaluate firms' corporate governance. We use Spearman pairwise correlations to assess convergent validity of raters' top-level scores, looking only at the top-level items pairs of raters have in common (Environmental, Social, Governance and Economic responsibility). As opposed to Pearson correlations, which assume scaled and ordered variables, Spearman pairwise correlations relax the scale assumption, which allow comparison between pairs of raters that do not use the same rating scale.

Insert Table 5 about here

In Table 5, the Spearman correlations between pairs of raters' top-level scores on their overlapping universes are fairly low. The average Spearman correlation of each rater ranges from -0.10 to 0.40. While KLD and Calvert environment ratings have reasonably high convergent validity, with a 0.63 correlation, Innovest environmental scores have low correlation with KLD scores (below

0.13). Asset4 environmental scores even have negative and statistically significant correlations with KLD (-0.23 in 2004, -0.11 in 2005 and -0.03 in 2006). Correlations between other high-level categories (Governance, Social, and Economic) are even lower. For instance, KLD Governance score are not significantly correlated with Asset4 and Innovest Governance scores. This additional evidence supports the idea that the low convergent validity between raters is not only due to different theorizations, but also to low commensurability. These findings were supported by several robustness tests. We first replicated results from Table 5 using our second measure of KLD top-level scores (Predictions from logit models instead of the sum of KLD strengths minus the sum of the concerns). Those results, presented in Appendix A7, also show low commensurability between raters, with KLD environmental score's correlation with other raters ranging from -.02 to .44, and the average Spearman correlation of the KLD governance score with other raters is 0.15.

Finally, in Table 6, we calculated the correlation over the 2002–2010 period between Asset4 and KLD data on low-level sub-scores (e.g., firms' involvement in "sin" industries, good relations with trade unions, or biodiversity protection). Table 6 highlights that reasonably high convergence occurs for some clearly defined sub-topics such as Tobacco involvement (0.63 correlation in 2010), but that a lack of commensurability still exists for more abstract subjects such as relations with trade unions or protection of indigenous people (respectively .15 and -.18 correlation in 2010). The prevalence of the latter kind of categories, where measurement is especially challenging, drives low convergent validity between these two SRI raters even after the adjustments discussed above.

Insert Table 6 about here

DISCUSSION

The primary contribution of this paper is to the literature on CSR, which increasingly utilizes the type of ratings used in our study. By finding little convergence among SRI ratings, our work is relevant for the hundreds of empirical studies on CSR that have used these data. Based on our study, we therefore urge scholars to test their empirical predictions using several ratings and be much more cautious about the conclusions derived from empirical work.

More specifically, our results should give pause to scholars who report high correlations of these ratings with outcomes (e.g. profits) and conclude that "doing good" and "doing well" are positively associated (this causal link has other challenges beyond the scope of this paper). The low convergent validity we report implies most or all of the metrics used in previous studies have low validity. Thus, our results shift the burden of proof to analysts using CSR ratings to show that the ratings are sufficiently valid for research purposes. Going forward, our results emphasize the importance of ongoing validation studies. Until such studies arrive that demonstrate a minimum level of convergent validity, authors should test if results can be replicated with multiple ratings. It is not enough to say one rating is best for a purpose. Unless the analyst can show a rating has sufficient validity, then it is best to use multiple measures to minimize problems of measurement error correlated with the predictor or outcome of interest.

Second, prior work has argued that CSR pressures across multiple dimensions in part explains the differential responses of firms, as managers seek to strategize the best way to deal with raters (Crilly et al., 2012; Delmas & Toffel, 2008; Philippe & Durand, 2011). However, it may be the lack of convergence among raters in the first place that actually accounts for the variation in responses across firms, an intriguing topic for future work.

Third, our findings are also relevant to a broader literature on ratings, pointing to an important boundary condition for prevailing theory. Prior work argued that raters distinguish themselves from one another on particular dimensions to establish a clear identity in the market (Negro, Hannan, & Rao, 2011). However, after accounting for distinct theorization, we fail to observe significant increases in convergent validity among raters. Raters' identity expressed in their theorization and methods does not explain rating divergence in our context. Hence, CSR ratings will have a limited impact on driving rated entities toward any particular shared behaviors.

In sum, as raters do not converge even when adjusted for differences in theorization, it is very likely that most of them also show a limited validity (i.e. they do not measure what they aim to measure), which is a serious concern not just for academics, but also for investors, activists, and policymakers. The market mediation as currently operated by SRI raters is unlikely to be socially optimal. Efforts to develop common measurement systems may lead to improvements in convergence. Indeed, recent consolidation in the SRI industry may force this convergence by merging several raters' theorizations and measures (e.g. MSCI now owns KLD and Innovest). We await future research to assess whether the next generation ratings are in fact increasing in validity.

REFERENCES

Chatterji, A. K., Levine, D. I., & Toffel, M. W. 2009. How Well Do Social Ratings Actually Measure Corporate Social Responsibility? ***Journal of Economics & Management Strategy***, 18: 125-169.

Crilly, D., Zollo, M., & Hansen, M. T. 2012. Faking it or muddling through? Understanding decoupling in response to stakeholder pressures. ***Academy of Management Journal***, 55(6): 1429-1448.

Delmas, M., Etzion, D., & Nairn-Birch, N. 2013. Triangulating environmental performance: What do corporate social responsibility ratings really capture? . ***The Academy of Management Perspectives***.

Delmas, M. A., & Toffel, M. W. 2008. Organizational responses to environmental demands: opening the black box. ***Strategic Management Journal***, 29(10): 1027-1055.

Devinney, T. M. 2009. Is the Socially Responsible Corporation a Myth? The Good, the Bad, and the Ugly of Corporate Social Responsibility. ***Academy of Management Perspectives***, 23(2): 44-56.

Durand, R., Rao, H., & Monin, P. 2007. Code and conduct in French cuisine: Impact of code changes on external evaluations. ***Strategic Management Journal***, 28(5): 455-472.

Entine, J. 2003. The Myth of Social Investing. ***Organization & Environment***, 16(3): 352.

Espeland, W. N., & Sauder, M. 2007. Rankings and Reactivity: How Public Measures Recreate Social Worlds. ***American Journal of Sociology***, 113(1): 1-40.

Galema, R., Plantinga, A., & Scholtens, B. 2008. The stocks at stake: Return and risk in socially responsible investment. ***Journal of Banking & Finance***, 32(12): 2646-2654.

Gray, R. 2010. Is accounting for sustainability actually accounting for sustainability…and how would we know? An exploration of narratives of organisations and the planet. ***Accounting, Organizations & Society***, 35(1): 47-62.

Hawken, P. 2004. Socially responsible investing. ***Sausalito, CA: Natural Capital Institute***.

Hsu, G., Roberts, P. W., & Swaminathan, A. 2012. Evaluative Schemas and the Mediating Role of Critics. ***Organization Science***, 23(1): 83-97.

Landis, J. Richard, and Gary G. Koch. "An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers." ***Biometrics*** 33.2 (1977): 363-74.

Mackey, A., Mackey, T. B., & Barney, J. B. 2007. Corporate Social Responsibility and Firm Performance: Investor Preferences and Corporate Strategies. ***Academy of management review***, 32(3): 817-835.

Margolis, J. D., & Walsh, J. P. 2003. Misery Loves Companies: Rethinking Social Initiatives by Business. ***Administrative Science Quarterly***, 48(2): 268-305.

Marquis, C., Glynn, M. A., & Davis, G. F. 2007. Community Isomorphism and Corporate Social Action. ***Academy of management review***, 32(3): 925-945.

McGahan, A. M., & Porter, M. E. 1997. How much does industry matter, really? ***Strategic Management Journal***(18 (1997)): pp. 15–30. 49.

Negro, G., Hannan, M. T., & Rao, H. 2011. Category Reinterpretation and Defection: Modernism and Tradition in Italian Winemaking. ***Organization Science***, 22(6): 1449-1463.

Orlitzky, M., Schmidt, F. L., & Rynes, S. L. 2003. Corporate Social and Financial Performance: A Meta-analysis. ***Organization Studies***, 24(3): 403-441.

Philippe, D., & Durand, R. 2011. The impact of norm-conforming behaviors on firm reputation. ***Strategic Management Journal***, 32(9): 969-993.

Rao, H., Monin, P., & Durand, R. 2003. Institutional Change in Toque Ville: Nouvelle Cuisine as an Identity Movement in French Gastronomy. ***American Journal of Sociology***, 108(4): 795-843.

Sauder, M., & Espeland, W. N. 2009. The Discipline of Rankings: Tight Coupling and Organizational Change. ***American Sociological Review***, 74(1): 63-82.

Sauder, M., & Wendy Nelson, E. 2009. The Discipline of Rankings: Tight Coupling and Organizational Change. ***American Sociological Review***, 74(1): 63-82.

Sharfman, M. 1996. The Construct Validity of the Kinder, Lydenberg & Domini Social Performance Ratings Data. ***Journal of Business Ethics***, 15(3): 287-296.

Waddock, S. 2003. Myths and Realities of Social Investing. ***Organization & Environment***, 16(3): 369.

TABLES

Table 1: Indexes' methodology

Indexes	Use of screens	Industry normalizing of the continuous score
Asset4 style	No	No
Innovest & DJSI style	No	Yes
KLD style	Firms with military concerns, tobacco concerns, alcohol concerns, and nuclear power concerns are screened out of the indexes	No
Calvert style	Firms with military concerns, tobacco concerns, and alcohol concerns are screened out of the index	Yes
FTSE4Good style	Firms with military concerns, tobacco concerns, and nuclear power concerns are screened out of the index	Yes

Table 2: Overlaps between SRI raters' indexes when overlapping universes are considered

	2004			*2005*				*2006*			
	Also in KLD DS400	Also in DJSI	Also in Asset4 A+	Also in KLD DS400	Also in Calvert	Also in Innovest	Also in Asset4 A+	Also in KLD DS400	Also in FTSE4Good	Also in Asset4 A+	Average overlap
KLD DS400		10%	16%		75%	3%	17%		24%	17%	29%
Calvert				41%		4%	12%				19%
Innovest				44%	59%		76%				60%
FTSE4Good								66%		39%	39%
DJSI	48%		40%								44%
Asset4 A+	54%	36%		47%	46%	16%		51%	43%		42%

Table 3: Pairwise tetrachoric correlations / Convergent validity of SRI raters on overlapping universes

	2004			*2005*				*2006*			
	KLD DS400	DJSI	Asset4 A+	KLD DS400	Calvert	Innovest	Asset4 A+	KLD DS400	FTSE4Good	Asset4 A+	Average correlation of this index
KLD DS400		0.45* N=2608	0.27* N=551		0.44* N=1072	- 0.00 N=555	0.12 N=631		0.40* N=629	0.16 N=615	0.26
Calvert				0.44* N=1072		0.07 N=508	- 0.12 N=617				0.13
Innovest				- 0.00 N=555	0.07 N=508		0.67* N=441				0.25
FTSE4Good								0.40* N=629		0.53* N=565	0.47
DJSI	0.45* N=2608		0.58* N=564								0.52
Asset4 A+	0.27* N=551	0.58* N=564		0.12 N=631	- 0.12 N=617	0.67* N=441		0.16 N=615	0.53* N=565		0.32
N = Universe * *p*-value <0.05								Average Correlation, EU Raters:			0.53
								Average Correlation, US Raters:			0.45
								Average Correlation, all Raters:			0.30
								Average Correlation, US & EU:			0.31

Table 4: Indexes' gaps
Top panel: top row is Asset4 standardized scores of each index's members minus the Asset4 standardized scores of its non-members / Other rows correspond to convergent validity after adjusting for explicit differences in theorization (industry screening and normalizing)

	2004			*2005*				*2006*		
Gaps	KLD DS400	DJSI	Asset4 A+	KLD DS400	Calvert	Innovest	Asset4 A+	KLD DS400	FTSE4Good	Asset4 A+
Asset4 Style	0.29**	1.15***	1.91***	0.18*	-0.21**	1.21***	1.82***	0.26**	0.90***	1.80***
KLD Style:	0.68***		1.31***	0.58***			1.20***	0.68***		1.28***
Calvert Style:					0.08		1.22***			
FTSE Style:									1.28***	1.13***
Innovest & DJSI Style:		1.10***	1.70***			1.22***	1.66***			

*** p<0.001, ** p<0.01, * p<0.05, + p<0.10

Table 5: Pairwise spearman correlations between KLD, Calvert, DJSI, Innovest, and Asset4's top-level scores on overlapping universes (Using KLD strengths minus concerns per category)

	2004			*2005*				*2006*		
	KLD	DJSI	Asset4	KLD	Calvert	Innovest	Asset4	KLD	Asset4	Average
Environmental score										
KLD		-0.09 N = 81	- 0.23* N = 551		0.63* N = 98	0.13* N = 554	-0.11* N = 631		-0.03 N = 616	0.05
Calvert				0.63* N = 98		0.35* N = 92	0.23* N = 92			0.40
DJSI	-0.09 N = 81		0.52* N = 53							0.22
Innovest				0.13* N = 554	0.35* N = 92		0.38* N = 441			0.29
Asset 4	- 0.23* N = 551	0.52* N = 53		-0.11* N = 631	0.23* N = 92	0.38* N = 441		-0.03 N = 616		0.13
Governance score										
KLD			-0.07 N = 551			0.04 N = 555	0.06 N = 631		0.06 N = 616	0.02
Innovest				0.04 N = 555			0.34* N = 441			0.19
Asset 4	-0.07 N = 551			0.06 N = 631		0.34* N = 441		0.06 N = 616		0.10
Social score										
DJSI			0.26 N = 53							0.26
Innovest							0.34* N = 441			0.34
Asset 4		0.26 N = 53				0.34* N = 441				0.30
Economic score										
DJSI			- 0.10* N = 53							-0.10

N = Universe ; * *p*-value <0.05

Table 6: Pairwise spearman correlations between KLD and Asset4's raw data 2002–2010 on overlapping universes

.	Tobacco involvement	Nuclear involvement	Military involvement	Gambling involvement	Alcohol involvement	Indigenous people	Biodiversity issues	Trade union relations	Average
2002	0.35* N = 374		0.79* N = 374	0.40* N = 374	0.67* N = 374	0.02 N = 374		-0.01 N = 374	0.37
2003	0.51* N = 386		0.78* N = 386	0.50* N = 386	0.66* N = 386	0.02 N = 386		-0.01 N = 386	0.41
2004	0.65* N = 524		0.67* N = 524	0.44* N = 524	0.50* N = 524	0.01 N = 524		-0.01 N = 524	0.38
2005	0.56* N = 598		0.56* N = 598	0.48* N = 598	0.54* N = 598	0.01 N = 598		0.08* N = 598	0.37
2006	0.65* N = 608	0.57* N = 33	0.62* N = 608	0.75* N = 608	0.64* N = 608	0.01 N = 608		0.15* N = 608	0.48
2007	0.82* N = 626	0.81* N = 103	0.66* N = 626	0.61* N = 626	0.63* N = 626	0.01 N = 626		0.28* N = 626	0.54
2008	0.89* N = 802	0.91* N = 91	0.67* N = 802	0.69* N = 802	0.82* N = 802	0.01 N = 802		0.19* N = 802	0.60
2009	0.89* N = 915	0.87* N = 72	0.71* N = 915	0.69* N = 915	0.87* N = 915	0.00 N = 915		0.18* N = 915	0.60
2010	0.63* N = 839	0.85* N = 40	0.64* N = 839	0.71* N = 839	0.65* N = 839	-0.18 N = 43	0.27* N = 659	0.15* N = 213	0.46

N = Universe * *p*-value <0.05

APPENDIX

Method Description-Tetrachoric correlations

To understand the meaning of tetrachoric correlations, we assume a standard measurement model:

$$R_{ij} = b\ T_i + e_{ij} \qquad \text{where:}$$

R_{ij} is the unobserved continuous score measured by an SRI rater *j* of firm *i*'s true level of responsibility;

T_i is the unobserved (latent) true level of social responsibility of firm *i*;

b is a regression coefficient; and

e_{ij} captures rater *j*'s measurement error and idiosyncratic definitions of "social responsibility."

For most of our raters (excluding KLD and Asset4), we only observe the discrete measure M_{ij} - whether SRI rater *j* has firm *i* as a member of its index. This membership equals one when the unobserved continuous rating R_{ij} is above SRI rater *j*'s cutoff ($Cutoff_j$), zero otherwise:

$$M_{ij} = 1 \text{ if } R_{ij} > Cutoff_j, \text{ and } 0 \text{ otherwise.}$$

Variation in $Cutoff_j$ is driven by each rater's desired membership size or by a rater's view of an acceptable minimum value. Tetrachoric correlation is a maximum likelihood technique that estimates the correlation of two raters' unobserved continuous ratings R_{ij} when only M_{ij} is observed. This measure is a correlation adjusted for the dichotomous nature of the data and the cutoff level of each rater.

References

Drasgow F. Polychoric and polyserial correlations. In Kotz L, Johnson NL (Eds.), Encyclopedia of statistical sciences. Vol. 7 (pp. 69-74). New York: Wiley, 1988.

Olsson U. Maximum likelihood estimation of the polychoric correlation coefficient. Psychometrika, 1979, 44(4), 443-460.

Table A1: Summary statistics of memberships

Membership in SRI indexes	IN	OUT	Universe (N)
2004			
KLD DS400	382	2231	2613
DJSI	88	2921	3009
Asset4 A+	61	548	609
2005			
KLD DS400	399	2603	3002
Calvert	607	490	1097
Innovest	18	585	603
Asset4 A+	91	583	674
2006			
KLD DS400	395	2199	2594
FTSE4Good	101	613	714
Asset4 A+	88	584	672

Table A2: Selection of firms' membership to SRI social indexes

	2004				2005					2006			
Membership in SRI raters social index	KLD DS400 index	DJSI index	Asset4 A+ index	% of membership	KLD DS400 index	Calvert index	Innovest index	Asset4 A+ index	% of membership	KLD DS400 index	FTSE 4Good index	Asset4 A+ index	% of membership
Google	No	No	No	**0%**	No	Yes	NR	No	33%	No	No	No	**0%**
Nike	No	Yes	NR	50%	Yes	Yes	No	No	50%	Yes	Yes	Yes	**100%**
Procter& Gamble	Yes	Yes	Yes	**100%**	Yes	Yes	No	Yes	75%	Yes	Yes	Yes	**100%**
Coca-Cola	Yes	No	No	33%	Yes	No	No	Yes	50%	Yes	Yes	Yes	**100%**
PepsiCo	Yes	No	Yes	67%	Yes	No	Yes	Yes	75%	Yes	No	Yes	67%
Time Warner	Yes	Yes	No	67%	Yes	Yes	No	Yes	75%	Yes	No	No	33%
Wall Mart	No	No	No	**0%**	No	No	NR	No	**0%**	No	No	Yes	33%
AT&T	Yes	No	No	33%	Yes	Yes	Yes	Yes	**100%**	Yes	Yes	No	67%
UPS	Yes	Yes	Yes	**100%**	Yes	Yes	Yes	Yes	**100%**	Yes	Yes	Yes	**100%**
Microsoft	Yes	No	Yes	67%	Yes	Yes	No	Yes	75%	Yes	Yes	Yes	**100%**
Amer. Express	Yes	No	No	33%	Yes	Yes	No	No	50%	Yes	Yes	No	67%
Bank of America	No	No	No	**0%**	No	Yes	Yes	No	50%	No	Yes	No	33%
Goldman Sachs	No	Yes	No	33%	No	Yes	No	Yes	50%	No	Yes	Yes	67%
General Motors	No	No	Yes	33%	No	No	No	Yes	25%	No	No	No	**0%**
General Electric	No	Yes	No	33%	No	No	No	Yes	25%	No	No	Yes	33%
Valero Energy	No	No	No	**0%**	No	No	No	No	**0%**	No	No	No	**0%**
Alcoa	No	Yes	NR	50%	No	No	Yes	No	25%	No	No	Yes	33%
Dow Chemical	No	Yes	Yes	67%	No	No	No	Yes	25%	No	No	Yes	33%
Pfizer	No	Yes	No	33%	No	Yes	No	Yes	50%	No	Yes	Yes	67%

NR: Not Rated

Table A3: Summary statistics for additional indexes

Membership in social indexes 2003–2005	IN	OUT	Universe (N)
2004			
KLD BMS	1945	668	2613
2005			
KLD BMS	2210	792	3002
KLD LCS	668	312	980
2006			
KLD BMS	1878	716	2594

Table A4: Panel A: Pairwise tetrachoric correlations / Convergent validity of SRI raters on overlapping universes

		KLD BMS	KLD LCS
2004	DJSI	- 0.12 N = 2613	
	Asset4 A+	- 0.16 N = 551	
2005	Calvert	0.69* N = 1072	0.69* N = 980
	Innovest	- 0.25 N = 555	- 0.23 N = 497
	Asset4 A+	- 0.27 N = 631	- 0.26* N = 609
2006	FTSE4Good	0.10 N = 629	
	Asset4 A+	- 0.09 N = 615	

N = Universe
* *p*-value <0.05

Panel B: 2003-2006 Pairwise tetrachoric correlations between Asset4 A+ and KLD DS400 on overlapping universes

	Asset4 A+ / KLD DS400
2003	0.08 N = 385
2004	0.26* N = 523
2005	0.08 N = 598
2006	0.14 N = 605

N = Universe
* *p*-value <0.05

Table A5: Pairwise tetrachoric correlations / Convergent validity of SRI raters for firms common to all raters' universes (551 in 2004, 413 in 2005, 538 in2006)

	2004				*2005*						*2006*				
	KLD BMS	KLD DS400	DJSI	Asset4 A+	KLD BMS	KLD LCS	KLD DS400	Calvert	Innovest	Asset4 A+	KLD BMS	KLD DS400	FTSE4Good	Asset4 A+	Average correlation of this index**
KLD BMS		1.00* N=551	0.03 N=551	-0.16 N=551		1.00* N=413	1.00* N=413	0.77* N=413	- 0.21 N=413	- 0.28* N=413		0.78* N=538	0.14 N=538	- 0.10 N=538	0.12
KLD LCS					1.00* N=413		1.00* N=413	0.77* N=413	- 0.21 N=413	- 0.28* N=413					0.09
KLD DS400	1.00* N=551		0.27* N=551	0.27* N=551	1.00* N=413	1.00* N=41		0.66* N=413	0.01 N=413	0.00 N=413	0.78* N=538		0.39* N=538	0.12 N=538	0.31
Calvert					0.77* N=413	0.77* N=413	0.66* N=413		0.10 N=413	- 0.12 N=413					0.44
Innovest					- 0.21 N=413	- 0.21 N=41	0.01 N=413	0.10 N=413		0.70* N=413					0.08
FTSE4Good											0.14 N=538	0.39* N=538		0.54* N=538	0.36
DJSI	0.03 N=551	0.27* N=551		0.58* N=551											0.29
Asset4 A+	-0.16 N=551	0.27* N=551	0.58* N=55		- 0.28* N=413	- 0.28* N=413	0.00 N=41	- 0.12 N=413	0.70* N=413		- 0.10 N=538	0.12 N=538	0.54* N=538		0.12
											Average Correlation, EU Raters:				0.54
											Average Correlation, US Raters:				0.47
											Average Correlation, all Raters:				0.29
											Average Correlation, US & EU Raters:				0.30

N = Universe

* *p*-value <0.05

** For KLD indexes only mean correlation with non-KLD indexes / For non-KLD indexes only mean correlation with KLD DS400

Table A6: Indexes' gaps
Top row is KLD standardized scores of each index's members minus the KLD standardized scores of its non-members / Other rows correspond to convergent validity after adjusting for explicit differences in theorization (industry screening and normalizing)

	2004			*2005*				*2006*		
Gaps	KLD DS400	DJSI	Asset4 A+	KLD DS400	Calvert	Innovest	Asset4 A+	KLD DS400	FTSE4Good	Asset4 A+
KLD Style:	1.02***	-0.27+	0.08	1.01***	1.27***	0.47	0.32	1.05***	1.48***	0.52*
Asset4 Style	0.77***		0.78***	0.81***			1.12***	0.86***		1.17***
Calvert Style:				0.98***	0.89***					
FTSE Style:								1.12***	1.45***	
Innovest & DJSI Style:	0.80***	0.89***		0.85***		2.20***				

*** p<0.001, ** p<0.01, * p<0.05, + p<0.10

Table A6 bis: Indexes' gaps
Top row is KLD standardized probability of inclusion in DS400 of index's members minus the KLD standardized probability of inclusion in DS400 of non-members / Other rows corresponds to convergent validity after adjusting for explicit differences in theorization (industry screening and normalizing)

	2004			*2005*				*2006*		
Gaps	KLD DS400	DJSI	Asset4 A+	KLD DS400	Calvert	Innovest	Asset4 A+	KLD DS400	FTSE4Good	Asset4 A+
KLD Style:	1.56***	1.63***	1.07***	1.45***	0.58***	1.17**	1.26***	1.42***	1.53***	1.35***
Asset4 Style	1.52***		1.41***	1.43***			1.83***	1.40***		1.66***
Calvert Style:				1.43***	0.51***					
FTSE Style:								1.44***	1.63***	
Innovest & DJSI Style:	1.49***	2.05***		1.40***		1.94***				

*** p<0.001, ** p<0.01, * p<0.05, + p<0.10

Table A7: Pairwise spearman correlations between KLD and other raters top-level scores on overlapping universes (Using probability of inclusion in DS400)

	2004		*2005*			*2006*	
	DJSI	Asset4	Calvert	Innovest	Asset4	Asset4	Average correlation
Environmental score							
KLD	0.29* N = 81	-0.02 N = 551	0.44* N = 98	0.24* N = 554	0.13* N = 631	0.23* N = 616	0.22
Governance score							
KLD		0.07 N = 551		0.24* N = 555	0.18* N = 631	0.12* N = 616	0.15

N = Universe
* *p*-value <0.05