Strategic Management Journal
Strat. Mgmt. J. (2015)
Published online EarlyView in Wiley Online Library (wileyonlinelibrary.com) DOI: 10.1002/smj.2407
Received 19 August 2013; *Final revision received 9 October 2014*

DO RATINGS OF FIRMS CONVERGE? IMPLICATIONS FOR MANAGERS, INVESTORS AND STRATEGY RESEARCHERS

AARON K. CHATTERJI,[1]* RODOLPHE DURAND,[2] DAVID I. LEVINE,[3] and SAMUEL TOUBOUL[4]
[1] *Duke University, Fuqua School of Business, Durham, North Carolina, U.S.A.*
[2] *HEC Paris, Center for Research on Society and Organizations, France*
[3] *Haas School of Business, University of California, Berkeley California, U.S.A.*
[4] *IPAG Business School, Department of Corporate Strategy & Finance, Paris, France*

Research summary: *Raters of firms play an important role in assessing domains ranging from sustainability to corporate governance to best places to work. Managers, investors, and scholars increasingly rely on these ratings to make strategic decisions, invest trillions of dollars in capital, and study corporate social responsibility (CSR), guided by the implicit assumption that the ratings are valid. We document the surprising lack of agreement across social ratings from six well-established raters. These differences remain even when we adjust for explicit differences in the definition of CSR held by different raters, implying the ratings have low validity. Our results suggest that users of social ratings should exercise caution in interpreting their connection to actual CSR and that raters should conduct regular evaluations of their ratings.*

Managerial summary: *Ratings of corporate social responsibility (CSR) guide trillions of dollars of investment, but managers, investors, and researchers know little about whether these ratings accurately measure CSR. In practice, there are examples of highly rated firms becoming embroiled in scandals and the same firm receiving sharply different ratings from different rating agencies. We evaluate six of the leading raters and find little overlap in their assessments of CSR. This lack of consensus suggests that social responsibility is challenging to measure reliably and that users of these ratings should be cautious in drawing conclusions about firms based on this data. We encourage the rating agencies to regularly validate their data in an effort to improve the measurement of CSR.*

INTRODUCTION

How much do we really know about corporate social responsibility (CSR)? Though many managers, investors, and scholars have embraced this concept, the ratings most often used to assess CSR have rarely been evaluated. If these ratings are invalid, then trillions of dollars of capital is potentially being misallocated and numerous academic findings may also not be valid.

In this study, we assess the convergent validity (that is, agreement) of six well-established social ratings. We find that these raters exhibit low convergence in their assessments of CSR.[1] This lack of agreement is not just due to announced differences in raters' theorization of CSR; for example, if they measure performance relative to an industry

Keywords: corporate social responsibility; ratings; corporate governance; socially responsible investing; performance measurement
*Correspondence to: Aaron K. Chatterji, Fuqua School of Business, Duke University, 1 Towerview Road, Durham, NC 27708, U.S.A. E-mail: ronnie@duke.edu

[1] When discussing the behavior of raters, we use the term *convergence*. When referring to the rating they provide, we use the term *convergent validity*. We do not wish to imply that convergence implies a particular time trend. We apply this term to describe overlap across ratings systems at a particular point in time.

group or in absolute terms. Instead, the low agreement implies all or almost all of the ratings have low validity. This result has important implications for managers, investors, and researchers who use these ratings.

Many managers spend significant time and resources on CSR activities. For example, analysts claim that nearly every Fortune 500 company releases some kind of sustainability report.[2] Eight thousand firms have signed the UN Global Compact as a sign of their commitment to CSR.[3] As CEOs and other top managers respond to growing pressure from multiple stakeholders over social issues (Bansal and Roth, 2000; Crilly, Zollo, and Hansen, 2012), high-profile and publically disseminated social responsibility ratings take on even greater importance. But if the ratings are not actually valid and cannot consistently identify socially responsible firms, then the hypothesized benefits of CSR cannot occur. For example, if managers cannot deduce whether their low rating is due to poor operations and performance, a different conceptualization of CSR than the raters, or simply poor measurement (Gray, 2010; Margolis and Walsh, 2003), then they will be unable to craft the appropriate response. 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.

Similarly, investors face serious challenges if metrics are invalid. If the enormous amount of socially responsible investment (SRI), approximately one out of every nine dollars in the United States[4] and one out of every six dollars in Europe (Cortez, Silva, and Areal, 2012), is being erroneously allocated to firms, then it implies significant inefficiencies in global capital markets. If 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 (Delmas, Etzion, and Nairn-Birch, 2013; Entine, 2003; Hawken, 2004), then SRI will not direct capital toward the most responsible firms. Thus, low convergent validity ensures the promise of "doing good and doing well" will be unfulfilled.

Academics should also be concerned about the convergent validity of SRI ratings. The academy has produced scores of articles on CSR and SRI over the past two decades (Orlitzky, Schmidt, and Rynes, 2003), with growing interest in recent years. For example, from 1994 to 2008, seven articles published in *SMJ* relied on data from just one of our SRI raters (KLD). From 2009 to 2013, 19 articles used KLD data and six articles employed other ratings we examine (FTSE4Good, Innovest, DJSI or Asset4). Notably, influential research has examined the effects of CSR on returns for investors and the cost of capital for managers (Galema, Plantinga, and Scholtens, 2008; Waddock, 2003). Other research has explored the drivers of CSR, such as profit-maximizing responses to heterogeneous consumer preferences (Mackey, Mackey, and Barney, 2007), imitation among firms, or a departure from profit-maximizing behavior to satisfy managers' private goals (Devinney, 2009; Marquis, Glynn, and Davis, 2007).

However, despite this growing interest in CSR, little research examines whether raters measure CSR accurately (Delmas *et al.,* 2013; Sharfman, 1996). If these metrics are invalid or are inconsistently applied across raters, then scholars who conduct analysis using one rating scheme risk drawing conclusions that are not accurate. Moreover, if there is systematic measurement error in SRI ratings, then scholars may report effects, for example, the positive or negative effect of CSR on firm performance, that are not true.

In short, it is crucial for managers, investors and academics to know the validity of social ratings and understand the dynamics driving convergence across raters. In this article, we first document that the ratings of six major social raters—KLD, Asset4, Calvert, FTSE4Good, DJSI, and Innovest—have fairly low correlations with each other. We then show that the correlation does not systematically increase when we adjust for announced differences in raters' theorization of CSR. Our results imply that SRI raters not only do not agree on one definition of responsibility (their "theorizations" of CSR differ), but also that raters may measure the same construct in different ways (the "commensurability" of CSR is low). Our findings suggest that consumers of this data should interpret SRI ratings with caution and validate these ratings before drawing strong conclusions about CSR.

[2] Kanani (2012).

[3] "From Fringe to Mainstream: Companies Integrate CSR Initiatives into Everyday Business." (http://knowledge.wharton.upenn.edu/article/from-fringe-to-mainstream-companies-integrate-csr-initiatives-into-everyday-business/) Last accessed July 21, 2013.

[4] US SIF Foundation, 2012 Report on Sustainable and Responsible Investing Trends in the United States.

DOI: 10.1002/smj

APPROACHING CONVERGENCE

The literature on social evaluations of firms and organizations establishes that two preconditions for convergence of raters. First, "theorization" makes clear precisely what raters assess and why it matters (Durand, Rao, and Monin, 2007; Hsu, Roberts, and Swaminathan, 2012). Next, "commensurability" of ratings makes comparison across raters possible (Espeland and Sauder, 2007; Sauder and Espeland, 2009).

"Theorization," according to Rao, Monin, and Durand (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, rated organizations can adjust their behaviors—or choose not to.

We use the term *theorization* to refer to the beliefs raters 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, the measurement and interpretation of key constructs are broadly similar across various financial rating agencies. We use the term *commensurability* to refer to the extent that raters get similar answers when they measure 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 the overlapping portions of what they define as *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 ("common 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."[5] KLD asserts that its "research is designed for investors and money managers who integrate environmental, social, and governance factors into their investment process."[6] 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 … ,"[7] 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."[8] In addition, all of the indexes cover similar high-level topics, including environmental and social performance.

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 financial metrics, while other raters do not. KLD, Asset4, FTSE4Good, and

[5] While our empirical analysis utilizes data from 2002 to 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 1, 2012).

[6] KLD's Research Products http://www.kld.com/research/index.html (Last accessed August 13th, 2007).

[7] Calvert-About the Ratings http://www.calvert.com/sri-index.html (Last accessed March 1st, 2012).

[8] Asset4 ESG content overview http://thomsonreuters.com/products_services/financial/content_news/content_overview/content_az/content_esg/ (Last accessed February 8th, 2012).

DOI: 10.1002/smj

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

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 percent of its subcategories[9] in the social issues domain. KLD therefore puts more weight on social issues than Asset4, a European rater, which has only 47 percent of its sub categories[10] related to social issues. In other domains, such as in issues relating to employees, Asset4 places more emphasis than 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 among employees' health and safety, training programs, and labor relations. KLD includes all of those topics under the subcategory 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 such as 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 four raters assert that CSR performance must be measured relative to industry peers (see Table 1).

The upshot is that, despite similar language, there are differences in the way various raters theorize 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 *et al.,* 2012; Delmas and Toffel, 2008; Philippe and 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 appeal to different groups.

However, low convergent validity will still be present in the case of low commensurability across raters; that is, when ratings of the same construct disagree due to differences in measurement. Thus, if we adjust for different theorizations (what constructs raters measure), then the convergent validity of ratings will be determined by differences in commensurability (how raters measure the same constructs). Commensurability is inherently a serious

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

[10] 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.

DOI: 10.1002/smj

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, and 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, such as 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, then adjusting for different theorizations should substantially increase convergent validity. For example, if all raters measure environmental performance using the same approach, then 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, then 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.[11] 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 United States; Innovest, in Canada; while FTSE4Good and Asset4 have origins in the Europe.[12] 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.

[11] 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.

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

DOI: 10.1002/smj

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&P 500, Russell 1000). Our first task was to denote the firms that had been included on each rater's index of top 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 that 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 3,134 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 Appendix 1 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 subscores, which rated each company on more specific aspects of their environmental and social performance. The KLD suscores 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 seven categories (Environment, Community, etc.).[13] We used these subscores 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 subscores.[14]

We had fewer details on other raters' subscores. For Calvert, we had five high-level scores,[15] but only for the 100 largest firms they rate. For DJSI, we had scores for its three high-level categories and for 78 firms that represented the within-industry top 10 percent 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 1–7 score for our analysis.

METHODS AND RESULTS

We first explore overlap among raters in terms of their assessments of CSR. In the Appendix 1, Table A2 shows that several well-known firms are included in some raters' social indexes, but not in the others. Google, for example, was considered as socially responsible only by Calvert in 2005. However, does this indicate that Google is not socially responsible? Or alternatively, that Google's CSR activity fits well with Calvert's theory of good CSR? Or that Calvert measures CSR in a way that erroneously advantages Google?

Table A2 provides initial insights about the low convergence of SRI raters. Strikingly, in 2004 at least six companies[16] are either in all or none of the most popular SRI raters' indexes.

We also explore convergence by measuring the likelihood that a company included in one index of responsible companies is also included in other indices. In doing this exercise, we must take into account that the raters' universes differ: for example, KLD only rates firms based in the United States. 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 percent.

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

[14] 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).

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

[16] UPS and Procter & Gamble are in all indexes. Walmart, Google, Valero Energy, and Bank of America are in none of the indexes.

DOI: 10.1002/smj

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

However, 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 1,000 and a second index includes only 10 firms from that universe, then no more than two percent of the first index can be members of the second index. Most common measures of agreement among binary ratings (e.g., the joint probability of agreement, the kappa statistics, and the Pearson and Spearman correlations) do not account for different memberships (and implicitly, for different s of what level of social responsibility is "enough" to be included in the index).

Second, 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 among 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 for the potentially distinct cutoff level of each rater (see Appendix 1 for further details). Importantly, tetrachoric correlations estimate the quantitative magnitude of the relationship between two raters in a fashion that is invariant to the number of companies selected in each index and that has familiar units (those of a Pearson correlation).

As an illustrative example, consider two psychiatrics who analyze the same population. Assume their assessment of patients' degree of depression is identical, but one perceives a much lower cutoff of when drugs are effective, so she prescribes drug therapy to far more patients. In such a case, the Pearson or Spearman correlations between treated and untreated patients will be low, while the tetrachoric correlation will score high.

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. The mean correlation is 0.30. That correlation implies that a firm that is 2 standard deviations high for one rater (i.e., a positive outlier in terms of

DOI: 10.1002/smj

social responsibility) is only 0.6 standard deviations high for the typical other rater (a bit above average).

Mean correlations between a given index and the other raters' indexes range from 0.13 (for Calvert) to 0.52 (for DJSI). Individual tetrachoric correlations between pairs of indexes range from −0.12 (between Calvert and Asset4 A+ in 2005) to 0.67 (between Innovest and Asset4 A+ in 2005). The several negative correlations indicate extreme disagreement: Firms that one rater considered socially responsible were *less* likely to be rated as responsible by the other rater than firms the first rater did *not* consider responsible. Only 3 of the 12 correlations are higher than 0.5.

However, while overall convergence is low, some similarities exist between groups of raters, specifically between raters based in the United States (KLD, DJSI, Calvert) and raters based in Europe (FTSE4Good, Asset4). The average tetrachoric correlations between U.S. raters (0.45) and between EU raters (0.53) are higher than the average correlation between all raters (0.31), providing suggestive evidence that geographically proximate raters may have closer theorizations and/or higher commensurability of CSR.

Correlations are similarly low when we examine other KLD indexes such as KLD BMS or KLD LCS (see Tables A3 and A4), and when we examine only the subgroup of firms that are common to every rater's universe (see Table A6). We also explore the tetrachoric correlations between KLD DS400 and Asset4 A+ over time on their overlapping universe of firms: 0.08 (2003), 0.26 (2004), 0.08 (2005), and 0.14 (2006). These results provide no evidence that convergent validity is improving (see Table A5).

There is no established cutoff that we are aware of to determine a "high" or "low" tetrachoric correlation. If the underlying data are normally distributed, then we can interpret tetrachoric correlations as we would Pearson correlations. For example, Kendler *et al.* (1992) describes a tetrachoric correlation of 0.68 as "quite strong" and 0.45 as "still substantial." Blanz, Schmidt, and Esser (1991) call 0.51 "moderate," and Thapar *et al.* (2000) label 0.4 as "relatively low." These descriptions appear analogous to the way strategy scholars think about Pearson correlations in our own research: 0.8 and above is generally thought of as "very high," and below 0.3 is usually described as "very low."

By this rule of thumb, agreement between SRI raters is low, especially when compared to related phenomenon in strategic management. For example, Dess and Robinson (1984) find high correlations across projections of future earnings and return on assets by managers in the same firm, ranging from 0.84 to 0.87. In their survey of management practices, Bloom and Van Reenen (2006) resurvey part of their sample and report a correlation of 0.73 with original assessments. It is crucial to appraise these possible benchmarks with regards to their respective settings. For example, one might expect ratings by managers in the same firm to have high agreement, while highly subjective domains such as movie ratings may lie at the other end of the spectrum. While there is a subjective component to social performance, each of the raters we study lists fairly specific criteria for assessment. Thus, we believe that the Bloom and Van Reenen (2006) management practice ratings are an appropriate available benchmark for assessing our results.

Taken together, the low tetrachoric correlations among the six raters, and the lack of improvement over time between KLD DS400 and Asset4 A+ implies there is low 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 nonmembers. 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 nonmembers. 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 nonmembers.

Our methodology follows this rationale. We first standardize Asset4 continuous scores ($R_{i\text{Asset4}}$) so that they have a 0 mean and a standard deviation of 1. We then compute the difference in the means of Asset4 continuous scores between members and nonmembers of each of the six indexes. Those

DOI: 10.1002/smj

Table 3. Pairwise tetrachoric correlations/convergent validity of SRI ratings on overlapping universes

	2004			2005				2006			Average correlation of this index
	KLD DS400	DJSI	Asset4 A+	KLD DS400	Calvert	Innovest	Asset4 A+	KLD DS400	FTSE4Good	Asset4 A+	
KLD DS400		0.45* N = 2,608	0.27* N = 551		0.44* N = 1,072	−0.00 N = 555	0.12 N = 631		0.40* N = 629	0.16 N = 615	0.26
Calvert				0.44* N = 1,072		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 = 2,608		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
								Average correlation, EU raters:			0.53
								Average correlation, U.S. raters:			0.45
								Average correlation, all raters:			0.30
								Average correlation, U.S. & EU:			0.31

*p-value <0.05.
N = Universe.

"membership gaps" are computed for each index i as follow:

$$\text{MembershipGap}_i = \frac{\sum_{c \text{ in index } i} s_c}{m} - \frac{\sum_{c \text{ not in index } i} s_c}{n - m}$$

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; and
- S_c is the standardization of R_c, the Asset4 score for company c.

The top row of the top panel of Table 4 shows the gaps in Asset4 scores of members and nonmembers of the other indices. They measure whether membership in one of the SRI indexes is a good predictor of the Asset4 continuous score. If raters had the same theorization and high commensurability, then these gaps should have similar values. However, while the gap between Asset4 Index members and nonmembers equals 1.80 standard deviations in 2006, for this same year, the gap between members and nonmembers of the FTSE4Good index is only 0.90 standard deviation and 0.26 for KLD-DS400. Members of the Calvert index even have Asset4 continuous scores significantly below the nonmembers (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. Overall, the gap in Asset4 scores between members and nonmembers averages 29 percent of the maximum possible gap.

Next, we adjust these gaps for differences in industry normalizing and screening.[17] The four lower rows of Table 4 present results when the gap in Asset4 continuous scores is recalculated using the screens and industry normalization of the specific other index. For example, in the second row we adjust Asset4 scores using KLD's screens.

In most cases, the gap between members and nonmembers increases and get closer to the recalculated gap for Asset4. 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. Overall, the gaps adjusting for explicit differences in theorizations close less than half the gap identified in the first row; the mean ratio of adjusted gaps/Asset4 gaps = 0.59.

[17] 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.

DOI: 10.1002/smj

Table 4. Indexes' gaps

	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***			

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$.
Top panel: top row is Asset4 standardized scores of each index's members minus the Asset4 standardized scores of its nonmembers/other rows correspond to convergent validity after adjusting for explicit differences in theorization (industry screening and normalizing).

Overall, these results provide evidence that different theorizations are responsible for part of the low convergent validity between raters. At the same time, 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 Table A7).

The third condition that explains divergences in rating is based on the nonoverlapping 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.

In Table 5, the Spearman correlations between pairs of raters' top-level scores on their overlapping universes are fairly low. Overall, the grand average Spearman correlation is 0.21.

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 convergence 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 aggregate 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 Table A8, also show low commensurability among raters. KLD environmental score's correlation with other raters ranges from -0.02 to 0.44, 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 eight low-level subscores (e.g., firms' involvement in "sin" industries, good relations with trade unions, and biodiversity protection). Table 6 highlights that reasonably high convergence occurs for some clearly defined subtopics 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

DOI: 10.1002/smj

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
Asset4	−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

*p-value <0.05.
N = Universe.

(respectively, 0.15 and −0.18 correlation in 2010). The prevalence of categories where measurement is challenging drives low convergent validity between these two SRI raters even after the adjustments discussed above.

DISCUSSION

The prior literature on raters argues that common theorization and commensurability are required for convergence. Across six sets of social ratings, we find limited evidence for common theorization, which can reduce convergent validity, but may still be consistent with high validity. Indeed, as long as users of each index understand the sources of divergence, divergent ratings can be valid measures of their own idiosyncratic definitions of *responsibility*.

However, we also find strong evidence of low commensurability of SRI ratings; that is, raters continue to have low agreement even when we adjust for explicit differences in what they say they are trying to measure. When commensurability is low, then all or most raters have high measurement error when trying to measure similar theoretical constructs. These results call into question the validity of social ratings, which impact managerial actions around the world, guide trillions of dollars of investment, and inform scholarly perspectives on corporate social responsibility.

We believe that these results should lead to careful assessments by managers, investors, and scholars as to what these ratings are capturing and how they should be used. If the ratings are invalid, then investors do not know which firms are the most responsible and risk misallocating trillions of dollars in capital. Further, managers lack clear

DOI: 10.1002/smj

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*		0.79*	0.40*	0.67*	0.02		−0.01	0.37
	N = 374		N = 374	N = 374	N = 374	N = 374		N = 374	
2003	0.51*		0.78*	0.50*	0.66*	0.02		−0.01	0.41
	N = 386		N = 386	N = 386	N = 386	N = 386		N = 386	
2004	0.65*		0.67*	0.44*	0.50*	0.01		−0.01	0.38
	N = 524		N = 524	N = 524	N = 524	N = 524		N = 524	
2005	0.56*		0.56*	0.48*	0.54*	0.01		0.08*	0.37
	N = 598		N = 598	N = 598	N = 598	N = 598		N = 598	
2006	0.65*	0.57*	0.62*	0.75*	0.64*	0.01		0.15*	0.48
	N = 608	N = 33	N = 608	N = 608	N = 608	N = 608		N = 608	
2007	0.82*	0.81*	0.66*	0.61*	0.63*	0.01		0.28*	0.54
	N = 626	N = 103	N = 626	N = 626	N = 626	N = 626		N = 626	
2008	0.89*	0.91*	0.67*	0.69*	0.82*	0.01		0.19*	0.60
	N = 802	N = 91	N = 802	N = 802	N = 802	N = 802		N = 802	
2009	0.89*	0.87*	0.71*	0.69*	0.87*	0.00		0.18*	0.60
	N = 915	N = 72	N = 915	N = 915	N = 915	N = 915		N = 915	
2010	0.63*	0.85*	0.64*	0.71*	0.65*	−0.18	0.27*	0.15*	0.46
	N = 839	N = 40	N = 839	N = 839	N = 839	N = 43	N = 659	N = 213	

*p-value <0.05.
N = Universe.

guidance in terms of which ratings to pay attention to, and scholars may derive influential conclusions about "doing good" and "doing well" that are not well-founded.

The low convergent validity we report implies that the results of prior academic studies using these metrics should be reassessed. Thus, we urge users to provide evidence that the ratings are sufficiently valid for their intended purposes. At minimum, for research purposes, it is best to use multiple measures as a robustness check to minimize problems of measurement error, especially error that may be correlated with the predictor or outcome of interest. We encourage researchers to acknowledge the error in social metrics, use statistical methods that adjust for measurement error, and/or justify why their chosen rating system is the right one to test their particular theoretical propositions.

We hope that our results will spur stakeholders who purchase these ratings to push social raters to validate their own ratings. Rather than implementing specific standards that might crowd out innovation, we would favor periodic assessments of these ratings using analyses similar to those employed in this article. Such validation can take many forms beyond the tests of convergent validity we present; for example (Chatterji and Toffel, 2010), testing whether environmental ratings correlate with objective measures such as harmful emissions and whether these ratings have predictive validity in terms of forecasting future environmental violations. Scholars can also perform additional studies; for example, testing whether highly rated firms have fewer major corporate scandals. Furthermore, scholars might undertake simulations to estimate precisely how much measurement error in social ratings affects empirical results in the academic literature. These simulations or similar analyses could also shed light on how much these measurement errors reduce expected returns and/or increase risk for socially conscious investors.

Finally, our work sheds light on two strands of scholarship on ratings. First, prior work has documented variation in responses by firms to the same ratings system. (Crilly *et al.,* 2012; Delmas and Toffel, 2008; Philippe and Durand, 2011). In our context, we have multiple raters, each with different theorizations of CSR, which could lead to even more heterogeneity in terms of how firms respond to ratings. Second, prior work argued that raters distinguish themselves from one another on particular dimensions to establish a clear identity in the market (Negro, Hannan, and Rao, 2011). However, after accounting for distinct theorization, we fail to observe much increase in convergent validity among raters. Rater identity, expressed in their published theorization and methods, does

DOI: 10.1002/smj

not explain divergence in our context, in contrast to more established fields (e.g., cuisine critics, wine tasters, financial analysts). In these contexts, clear (although debated) theorization and commensurability are preconditions for rated entities to converge to common behaviors. In our setting, there is not enough overlap among the raters themselves in terms of how to measure CSR to even begin this process of convergence. Hence, SRI ratings will have a limited impact on driving rated firms toward any particular shared behaviors, and the market mediation provided by SRI raters is unlikely to be socially optimal. Efforts to develop common measurement systems may lead to an improvement in convergence. Indeed, recent consolidation in the SRI industry may actually compel 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 increasing in validity.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support from the HEC Foundation, the GDF-Suez Chair on Business and Sustainability and the Society and Organizations (SnO) Research Center. We also thank our two anonymous reviewers for their guidance in the development of this paper.

REFERENCES

Bansal P, Roth K. 2000. Why companies go green: a model of ecological responsiveness. *Academy of Management Journal* **43**(4): 717–736.

Blanz B, Schmidt MH, Esser G. 1991. Familial adversities and child psychiatric disorders. *Journal of Child Psychology and Psychiatry* **32**(6): 939–950.

Bloom N, Van Reenen J. 2006. Measuring and explaining management practices across firms and countries. National Bureau of Economic Research Working paper No. w12216.

Chatterji AK, Levine DI, Toffel MW. 2009. How well do social ratings actually measure corporate social responsibility? *Journal of Economics and Management Strategy* **18**: 125–169.

Chatterji AK, Toffel MW. 2010. How firms respond to being rated. *Strategic Management Journal* **31**(9): 917–945.

Cortez MC, Silva F, Areal N. 2012. Socially responsible investing in the global market: the performance of US and European funds. *International Journal of Finance and Economics* **17**(3): 254–271.

Crilly D, Zollo M, Hansen MT. 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? *Academy of Management Perspectives* **27**(3): 255–267.

Delmas MA, Toffel MW. 2008. Organizational responses to environmental demands: opening the black box. *Strategic Management Journal* **29**(10): 1027–1055.

Dess GG, Robinson RB. 1984. Measuring organizational performance in the absence of objective measures: the case of the privately-held firm and conglomerate business unit. *Strategic Management Journal* **5**(3): 265–273.

Devinney TM. 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.

Drasgow F. Polychoric and polyserial correlations. In Kotz L, Johnson NL (Eds.), *Encyclopedia of Statistical Sciences* (Volume 7) pp. 69–74. New York: Wiley, 1988.

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 and Environment* **16**(3): 352–368.

Espeland WN, 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 and 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 and Society* **35**(1): 47–62.

Hawken P. 2004. *Socially Responsible Investing*. Natural Capital Institute: Sausalito, CA.

Hsu G, Roberts PW, Swaminathan A. 2012. Evaluative schemas and the mediating role of critics. *Organization Science* **23**(1): 83–97.

Kanani R. 2012. The future of CSR. Available at: http://www.forbes.com/sites/rahimkanani/2012/02/09/the-future-of-corporate-social-responsibility-csr/ (accessed 21 July 2013).

Kendler KS, Neale MC, Kessler RC, Heath AC, Eaves LJ. 1992. Major depression and generalized anxiety disorder: same genes, (partly) different environments? *Archives of General Psychiatry* **49**(9): 716–722.

Mackey A, Mackey TB, Barney JB. 2007. Corporate social responsibility and firm performance: investor preferences and corporate strategies. *Academy of Management Review* **32**(3): 817–835.

Margolis JD, Walsh JP. 2003. Misery loves companies: rethinking social initiatives by business. *Administrative Science Quarterly* **48**(2): 268–305.

DOI: 10.1002/smj

Marquis C, Glynn MA, Davis GF. 2007. Community isomorphism and corporate social action. *Academy of Management Review* **32**(3): 925–945.
Negro G, Hannan MT, Rao H. 2011. Category reinterpretation and defection: modernism and tradition in Italian winemaking. *Organization Science* **22**(6): 1449–1463.
Olsson U. 1979. Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika* **44**(4): 443–460.
Orlitzky M, Schmidt FL, Rynes SL. 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 WN. 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 and Domini social performance ratings data. *Journal of Business Ethics* **15**(3): 287–296.
Thapar A, Harrington R, Ross K, McGuffin P. 2000. Does the definition of ADHD affect heritability? *Journal of the American Academy of Child and Adolescent Psychiatry* **39**(12): 1528–1536.
Waddock S. 2003. Myths and realities of social investing. *Organization and Environment* **16**(3): 369–380.

APPENDIX 1: METHOD DESCRIPTION-TETRACHORIC CORRELATIONS

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

$$R_{ij} = b\, T_i + e_{ij}$$

where:

1. R_{ij} is the unobserved continuous score measured by an SRI rater j of firm i's true level of responsibility;
2. T_i is the unobserved (latent) true level of social responsibility of firm i;
3. b is a regression coefficient; and
4. 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 1 when the unobserved continuous rating R_{ij} is above SRI rater j's cutoff (Cutoff_j), 0 otherwise: $M_{ij} = 1$ *if* $R_{ij} > \text{Cutoff}_j$, and 0 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 (see Drasgow 1988 and Olsson 1979 for references).

Table A1. Summary statistics of memberships

Membership in SRI indexes	IN	OUT	Universe (N)
2004			
KLD DS400	382	2,231	2,613
DJSI	88	2,921	3,009
Asset4 A+	61	548	609
2005			
KLD DS400	399	2,603	3,002
Calvert	607	490	1,097
Innovest	18	585	603
Asset4 A+	91	583	674
2006			
KLD DS400	395	2,199	2,594
FTSE4Good	101	613	714
Asset4 A+	88	584	672

DOI: 10.1002/smj

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

Membership in SRI raters social index	2004				2005					2006			
	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	FTSE4Good 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%
Walmart	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%**
American 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.

DOI: 10.1002/smj

Table A3. Summary statistics for additional indexes

Membership in social indexes 2003–2005	IN	OUT	Universe (N)
2004			
KLD BMS	1,945	668	2,613
2005			
KLD BMS	2,210	792	3,002
KLD LCS	668	312	980
2006			
KLD BMS	1,878	716	2,594

Table A4. Pairwise tetrachoric correlations/convergent validity of SRI raters on overlapping universes

		KLD BMS	KLD LCS
2004	DJSI	−0.12	
		N = 2,613	
	Asset4 A+	−0.16	
		N = 551	
2005	Calvert	0.69*	0.69*
		N = 1,072	N = 980
	Innovest	−0.25	−0.23
		N = 555	N = 497
	Asset4 A+	−0.27	−0.26*
		N = 631	N = 609
2006	FTSE4Good	0.10	
		N = 629	
	Asset4 A+	−0.09	
		N = 615	

*p-value <0.05.
N = Universe.

Table A5. 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

*p-value <0.05.
N = Universe.

DOI: 10.1002/smj

Table A6. Pairwise tetrachoric correlations/convergent validity of SRI raters for firms common to all raters' universes (551 in 2004, 413 in 2005, 538 in 2006)

	2004				2005						2006				Average correlation of this index[a]
	KLD BMS	KLD DS400	DJSI	Asset4 A+	KLD BMS	KLD LCS	KLD DS400	Calvert	Innovest	Asset4 A+	KLD BMS	KLD DS400	FTSE4Good	Asset4 A+	
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, U.S. raters:	0.47
														Average correlation, all raters:	0.29
														Average correlation, U.S. & EU raters:	0.30

[a] For KLD indexes only mean correlation with non-KLD indexes/for non-KLD indexes only mean correlation with KLD DS400.
*p-value <0.05.
N = Universe.

Table A7. Indexes' gaps. (a) Top row is KLD standardized scores of each index's members minus the KLD standardized scores of its nonmembers/other rows correspond to convergent validity after adjusting for explicit differences in theorization (industry screening and normalizing). (b) Top row is KLD standardized probability of inclusion in DS400 of index's members minus the KLD standardized probability of inclusion in DS400 of nonmembers/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+
(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***				
(b)										
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 A8. Pairwise Spearman correlations between KLD and other raters top-level scores on overlapping universes (using probability of inclusion in DS400)

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

*p-value <0.05.
N = Universe.

DOI: 10.1002/smj