UC Berkeley
Working Paper Series

Title
Imitate or Differentiate? Evaluating the validity of corporate social responsibility ratings

Permalink
https://escholarship.org/uc/item/3sz7k7jc

Authors
Chatterji, Aaron K
Levine, David I. I.

Publication Date
2008-02-05

Peer reviewed

Imitate or Differentiate?
Evaluating the validity of corporate social responsibility ratings [*]

Aaron K. Chatterji
Fuqua School of Business
Duke University
1 Towerview Drive
Durham, NC 27708
ronnie@duke.edu

David I. Levine
Haas School of Business
University of California at Berkeley
545 Student Services Building #1900
Berkeley, CA 94720-1900
levine@haas.berkeley.edu

December 2007

Imitate or Differentiate?
Evaluating the validity of corporate social responsibility ratings

Abstract:

Although there is $2 trillion in portfolios using socially responsible investing (SRI) criteria, it remains unclear how to measure "social responsibility." We explore competing theoretical perspectives that explain the level of convergent and predictive validity across SRI ratings produced by competing social raters. While some prior literature predicts low convergent validity due to desire for differentiation, other work predicts high convergent validity driven by high true validity or by neo-institutionalist forces that reward imitation. We find that these ratings have low correlations and that firms with high and low social ratings are equally likely to be later embroiled in scandals.

In 2005, professional fund managers invested at least $2 trillion with some consideration of "corporate social responsibility" in mind.[1] The huge amount of capital under the banner of socially responsible investing (SRI) has drawn considerable attention from scholars, activists, managers, and policymakers. Some advocates of corporate social responsibility praise SRI, believing that it can direct capital towards the most responsible firms while penalizing firms with poor social performance. At the same time, skeptics argue that the organizations that rate the social performance of enterprises, referred to as "raters" in our study, cannot truly discern which

---

[1] Social Investment Forum 2005 Report. This socially conscious segment is almost 10% of the total of funds managed by professional investors. At the same time, many of these investors screen only for tobacco, while we consider funds with a broader set of criteria in this paper.

firms are socially responsible, resulting in metrics that are often invalid and can be misleading to stakeholders (Entine, 2003).

For their part, academics have produced dozens of articles on SRI (see recent review by Orlitsky, Schmidt, and Rynes, 2003). Most recent research has examined whether SRI affects returns for investors and the cost of capital for managers (see Waddock, 2003 for a review). A more fundamental question is whether commonly used indicators of social responsibility are valid measures of the social, environmental and ethical performance of enterprises. If these metrics are invalid, none of the hypothesized benefits of socially responsible investing can occur. In the worst-case scenario, if firms expend resources to achieve high scores on invalid metrics, then social welfare can decline when managers pay more attention to scoring highly on social ratings. Thus, it is crucial to understand the validity of the metrics used by SRI raters. Unfortunately, almost no careful validations of competing SRI metrics have been conducted. (For one example, see Sharfman, 1996)

Some scholars have noted particular problems with SRI ratings metrics. Paul Hawken (2004) points out that the various methodologies employed by socially responsible raters allow for almost any public firm to be considered for at least one SRI index. John Entine (2003) accuses raters of giving high marks to firms that are later more likely to be embroiled in scandals. Our paper expands these anecdotal critiques into formal tests of convergent validity (that is, whether social ratings agree with one another, after adjusting for purposeful differences) and of predictive validity (that is, whether social ratings predict later scandals).

Our tests are based on two broad sets of theories that try to explain why the social ratings of competing SRI firms will be similar and/or different.[2] The first set of theories predicts that social ratings from competing firms will be strongly correlated. This high correlation may be driven by either high-quality measurement of agreed-on standards or by neo-institutionalist forces pushing for imitation. The second set of theories predicts low correlation, either due to raters attempting to inform stakeholders with different preferences or due to measurement error. That measurement error, in turn, can either be random or purposeful as raters attempt to appear as if they have unique information (a form of product differentiation as fashion, discussed in Bourdieu [1984] and modeled formally in Zitzewitz [2001]).

However, neo-institutionalist forces for imitation can lead raters to agree on a socially constructed measure of "responsibility" even when that agreed-on measure is not valid. If such forces are strong we can have high convergent validity (that is, agreement among raters) without high validity (that is, correlation of social ratings and true corporate social performance).

Thus, we also examine the predictive validity of two of the ratings. To study *predictive validity* we chose an uncontroversial minimal level of social responsibility: Whether highly rated firms are less likely to be involved in a major scandal than other firms of similar size in their industry. In the case of high-quality measurement of agreed-on standards, but not correlated measurement error, we expect firms with scandals to have received below-average social ratings in previous years.

---

[2] See Baum and Haveman (1997) for an overview of these two sets of theories.

We found major social ratings to have a fairly low correlation with each other, supporting theories of differentiation. Such differences do not appear to be due to differences in measurement strategies that are valuable to investors, as the correlations do not systematically increase when we adjust for differences that investors can easily observe. These results of low convergent validity mean that all or most of the SRI ratings are not measuring "true" social responsibility. Because we make no claim to what that "true" measure might be, we cannot measure which (if any) of the metrics approaches that ideal.

We also found that firms with high and low social ratings are equally likely to be embroiled in a major scandal a few years later – although this test of predictive validity has low statistical power. Overall the results show limited validity, which is a serious concern for investors, academics, activists, and policymakers. In the next section, we expand on the theoretical work discussed above and provide a brief overview of the socially responsible investing sector. We then present our hypotheses. Subsequently, we discuss our datasets and methods, our results, and offer some concluding remarks.

Theoretical Perspectives on Convergent Validity

We first expand on the several theoretical perspectives discussed above. We then provide a brief overview of the socially responsible investing sector and explain how the theories apply in this empirical setting.

Imitation vs. Differentiation

Economists and sociologists have described economic incentives and social forces that can lead all kinds of firms to both differentiate products from those of their competitors and to imitate the

products of their competitors. For example, Baum and Haveman (1997) found that hoteliers have reasons to locate close to competitors (in imitation of competitors' locational strategies) and also to build smaller or larger establishments (to differentiate their product). Choosing a location close to existing hotels creates agglomeration benefits and can also increase competition. Hoteliers mitigate these competitive threats by building near hotels of different sizes. Thus, this study found support for theories that predict differentiation as well as those that predict imitation in competitive markets.

Theories of Convergence
In this section we describe two families of theories of convergence – in our context, of different social raters providing similar ratings to firms. Only the first of these two theories also implies high true validity.

Convergence with high validity. In the case of ratings, social raters may provide similar ratings due to multiple raters measuring the same construct (that is, definition of social responsibility) with high-quality measurement methods and data. Such convergence will lead to both high convergent validity and high true validity.

Convergence with low validity. Neo-institutional theory offers an alternative explanation for similarity that is especially applicable to a nascent industry such as socially responsible investing. The critical assumption in neo-institutional theory is that organizations seek legitimacy in their environment, rather than strictly efficiency (Scott, 1995; Staw and Epstein, 2000). Staw and Epstein make the point especially relevant to measurement of social responsibility:

> *When technologies are poorly understood and organizations face problems with ambiguous causes and unclear solutions, copying other organizations (and their executives) may simply be a low-cost heuristic for finding useful solutions (Staw and Epstein 2000, citing March and Olsen 1976).*

Pressures to imitate others may be especially strong when the underlying mechanisms are not clear to decision makers ("mimetic isomorphism" in the language of DiMaggio and Powell 1983) or when norms suggest certain behaviors are legitimate ("normative isomorphism").

Some of the relevant insights from neo-institutionalist theory have been formalized in the economics and finance literature on information cascades that lead to herding behavior. In these models, individuals and organizations have limited information and can thus learn about their environment in part by observing the decisions of their peers. The result can be herd-like behavior even when most decision-makers have information that would recommend a different outcome (Banerjee, 1992). Evidence for such herding behavior can be found among physicians making a diagnosis (Bikhchandani, Hirshleifer, and Welch, 1992), and voters choosing a candidate (Bartels, 1988). More closely related to the ratings we examine, Scharfstein and Stein (1990) find that investment managers' concern about their reputations can lead them to ignore their private information and follow the herd by putting their clients' money in assets *popular with other managers*. There is also evidence of such herding among securities analysts (Rao, Greve, and Davis, 2001; Hong, Kubik, and Solomon, 2000) and some credit raters (Vaaler and McNamara 2004).

These models of herding (Scharfstein and Stein 1990; Rao, Greve, and Davis, 2001; Hong, Kubik, and Solomon 2000; and Vaaler and McNamara 2004) focus on individual behavior, taking as given the behavior of the organizations they rate. Modern organizational theories emphasize that organizations often act to affect their environment. Meyer and Gupta (1994), for example, describe how managers who are rated often try to affect their ratings.

Theories of Differentiation

Differentiation with high validity. Theories of product differentiation will lead to low

convergent validity, but that low convergent validity need not imply low true validity when

sellers are appealing to diverse consumer preferences. For example, the best Chinese restaurant

in a town can have a vastly different menu than the best French bistro, even though they both

have excellent food. In the world of finance, an investment advisor may suggest a risky portfolio

for a wealthy young investor and a safer portfolio for a working-class retiree. That divergence of

advice is consistent with high validity of the advice in each case – given the different situations

of the two investors.

Differentiation with low validity. Alternatively, differences in ratings may be due to raters'

measurement error. Most obviously, measurement error can be fairly random due to arbitrary

differences in how raters define and measure the construct of interest.

Interestingly, differentiation can also arise for legitimation purposes. Bourdieu (1984) took a

sociological look at fashion, finding that designers differentiate their products not to increase

functionality, but largely for the sake of difference. (For a more recent statement, see Crane

1999) Zitzewitz (2001) provides an economic model of a subset of Bourdieu's theory, noting

that career concerns can give some analysts an incentive to exaggerate their differences from the

herd as a means of indicating they hold private information.

Applications to SRI

The literature discussed above explains the theoretical motivations for firms to imitate each other

or differentiate their product offerings. In this section we apply these abstract arguments to the

SRI sector, where raters' product is in fact their ratings. While there has been some theoretical

and empirical literature on the forces pushing analysts of firms' **financial** performance to converge or diverge (Zitzewitz, 2001), almost no prior work has examined how these forces operate among analysts of firms' **social** performance. We analyze data from five of the major social raters in order to do so: KLD, Calvert, FTSE4Good, DJSI, and Innovest. Appendix 1 describes these social raters in more detail.

Theories of convergence
Convergence with high validity. Convergence with high validity occurs if social raters correctly agree on what social responsibility looks like and know how to measure it. There is agreement at a high level of abstraction on what constitutes good social performance. Specifically, all of the indices cover similar high-level topics: environment, workplace, business practices, human rights, and community relations.[3]

While not all of the raters are explicit in what they measure, they make public claims (that is, that investors can see) which are quite similar as well as very general. For example, one of FTSE4Good's stated goals is "To provide a tool for responsible investors to identify and invest in companies that meet globally recognised 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."[5] Calvert describes its ratings as "a comprehensive system for analyzing and ranking company corporate responsibility policies and performance across five key areas…".[6]

---

[3]These are the five Calvert sub-scores, but all of the ratings we examine include items that fall within all five of these broad categories.
[4] FTSE4Good Inclusion Criteria
http://www.ftse.com/Indices/FTSE4Good_Index_Series/Downloads/FTSE4Good_Inclusion_Criteria_Brochure_Feb_06.pdf (Last accessed August 13th, 2007)
[5] KLD's Research Products http://www.kld.com/research/index.html (Last accessed August 13th, 2007)
[6] Calvert-About the Ratings http://www.calvert.com/sri_7894.html (Last accessed August 13th, 2007)

The ratings also all share face validity. That is, raters make a significant effort to persuade potential investors that their methods and ratings are based on careful analysis and do not simply imitate those of their competitors. For example, social raters draw on multiple sources and use multiple research methods, which are established scientific approaches: they review official government data (e.g., on toxic emissions and regulatory actions), company documents, press reports; they conduct interviews; they even deploy surveys. Their marketing literature stresses the care raters give to careful analysis of company's social record. Finally, they compare themselves to traditional financial research firms. For example, KLD describes its services as "analogous to those provided by financial research service firms." Innovest touts its expertise "with a particular focus on their impact on competitiveness, profitability, and share price performance." [7] In fact, Dow Jones and the *Financial Times* are also well known providers of traditional financial information.

Convergence with low validity. Convergence (and high convergent validity) does not necessarily mean that raters have high true validity, in large part because they leave the underlying construct of "social performance" somewhat vague.

This vagueness leaves room for mimetic and normative isomorphism to lead to common measurement error. Looking across firms, social raters may have converged on an erroneous definition of "social performance." For example, some theories in what is known as "deontological ethics" suggest it is evil for someone to profit from owning companies that

---

[7] For more examples, see KLD's "About Us" section at http://www.kld.com/about/index.html (Last accessed June 28th, 2007) "Methodology" section at http://www.kld.com/research/methodology.html (Last accessed June 28th, 2007), Innovest homepage at http://www.innovestgroup.com/ (Last accessed August 13th, 2007)

employ child labor regardless of the consequences of that ownership. If firms react to social ratings based on these ethical theories by dismissing child labor and the children, thus, are made worse off, a consequentialist might claim such social rating can make the world a worse place.[8]

In addition, consistent with the Meyer and Gupta (1994) point that those measured often try to manipulate ratings, rated companies have incentives to engage in "greenwashing," or highlighting sometimes modest achievements in their environmental and social performance (Lyon and Maxwell, 2005). All of the ratings firms emphasize that a significant share of their underlying data come from the rated companies themselves. Moreover, corporate public relations make major efforts to influence press reports (another main source of social raters' information). If some companies excel at misrepresenting their social performance, we have an additional explanation for correlated measurement error.

Theories of differentiation
While the raters are generally consistent in their philosophy and the outline of their measurement method, several important differences are visible to investors and other users of their ratings, as this section summarizes. For a more complete description of how the indices vary in construction, see Chatterji and Levine (2006).

First, all of the social raters except KLD rank firms within industry or sector. Next, three of the five raters we examine (KLD, Calvert and FTSE4Good, but not DJSI or Innovest) use explicit screens to rule out some firms. All those with screens screen out firms with substantial military and tobacco interests, although their precise definitions of "substantial" vary and FTSE4Good

---

[8] See Basu and Tzannatos (2003) for a survey of the relevant theory and evidence related to child labor.

only screens out nuclear weapons makers. KLD and Calvert screen out alcohol makers. KLD and FTSE4Good screen out firms with revenue from nuclear power.

The relative weights that each social rater puts on each top-level category (environment, human rights, etc.) also vary. KLD has no explicit weights on sub-scores, but a committee reviews all the evidence and sub-scores and decides which companies to include. Calvert chooses funds similarly, with 5 top-level categories and no explicit weights detailed on their website; DJSI places explicit weights of 1/3 each on economic, environmental, and social sub-scores. FTSE4Good has weights that vary by industry (for example, with more weight on the environment in environmentally sensitive sectors), while Innovest scores firms in 36 different categories and uses different weighting schemes in each industry.

The precise items within each top-level sub-score also vary across raters. At the same time, most investors would not be able to see such items or their decision weights for most of the ratings.[9] We had to negotiate and/or pay for access to company reports for some of the raters. Given these fees, and the many hours it has taken us to understand the measurement methods of each rater, we are confident that a meaningful share of socially-conscious investors do not take these subtleties into account.

Differentiation with high validity. It is possible that differences in ratings are due to raters' strategic choices to satisfy the demands of different groups of clients. In that case, low correlations across measures would be consistent with high validity of each metric in measuring its focal construct (that is, low convergent validity when we treat the ratings as measuring the same construct, but high validity when we relate each rating to its distinct construct). Intuitively,

---

[9] DJSI and FTSE4Good however, have their full weighting scheme on their websites

as discussed above, it is not a critique of a neighborhood's restaurant quality if it has restaurants with cuisines from many continents—even if their recipes are very different. Similarly, socially-conscious investors might have multiple motives (outlined in appendix 2). This variation in investors' social preferences should lead to a variety of social portfolios with different measurement strategies and, thus, different companies in their portfolio.[10]

For example, when social raters use screens against tobacco or other categories of firms, that screen is useful to investors who wish to avoid profits from these sectors ("deontological" and "expressive" investors in the language of appendix 2). Yet there is no consistent theory or evidence that investors trying to achieve excess returns should also avoid such sectors.

In contrast, Innovest, one of the newest social raters we examine, has no social screens. Its rhetoric largely reflects the search for measures of good management that will lead to high financial returns[11]. If investors using these data understand the measurement practices and principles underlying the data, then any divergence in ratings can be valuable in satisfying diverse investor types.

Similarly, four of the five raters we examine measure social performance relative to each industry, while KLD does not do so. This practice is also visible to investors and the heterogeneity may be useful if investors differ on whether they do or do not prefer measurement relative to the rest of each industry.

---

[10] Entine (2003) discusses the variation in screens directed at investors with different social preferences as a critique of SRI. We disagree and consider it a feature of market systems that people with different preferences can make different choices – as long as those making the choices are aware of the differences in screens and other measurement strategies.

[11] Innovest's website states "At Innovest, we think about investing a bit differently. For us, companies' ability to handle political, environmental, labour, and human rights risks are powerful proxies and leading indicators for their overall management quality – or the lack thereof."
http://www.innovestgroup.com/index.php?option=com_content&task=blogcategory&id=24&Itemid=32 (Last accessed August 13th, 2007)

But if ratings and portfolios vary substantially after we adjust ratings for the visible differences in measurement strategy, then it is likely that all or almost all of the portfolios have high measurement error. An issue recurring in the discussion below is the importance of evaluating the extent to which differences in SRI ratings are to desirable variation in measuring explicitly different constructs and the extent to which it is due to undesirable measurement error of very similar constructs.

Differentiation with low validity. The raters' overall agreement on the broad outlines of social responsibility nonetheless leaves room for substantial measurement error. At a basic level, raters are unclear about what values underlie social performance. For example, as noted above, some deontological theories suggest it is evil to support firms whose suppliers employ child labor, while consequentialist theories suggest it is immoral to support firms that penalize (for example, by firing them) children just because they are so poor that they must work for pay. Raters are typically unclear about what weights they should use to aggregate these ethical approaches. Similarly, there are no guidelines for how to trade off progress on global warming versus a poor record on discrimination. Raters are also unclear how to how to measure the various components of social performance, in large part because they do not know how observable corporate behavior causally effects social outcomes. For example, a social rater interested in measuring whether a company provides equal opportunity to under-represented minorities may be limited to a proxy such as the minority membership on its board of directors. (This specific measure is used by at least 1 of the 5 social raters we examine.) No evidence exists correlating that important construct with that easy-to-observe proxy.

In a world where investors, companies, and the raters themselves are uncertain about what domains of social responsibility are important and how to measure each domain, raters may have incentives to deviate from the crowd to differentiate their product and to send a signal of superior ability (Zitzewitz 2001). Thus, social raters may include unusual firms in their "approved" indexes or rate firms much higher or lower than their competitors. Thus, we may see anti-herding behavior in some cases. Regardless of the intentionality, either form of measurement error implies that low convergent validity will imply low validity.

Hypotheses

We first present our hypotheses related to convergent validity and then for predictive validity.

Convergent validity

Two metrics that attempt to measure the same construct have convergent validity if they correlate well with each other. For example, consider the SAT test for college admissions. If the math and reading portion of the SAT do not do not correlate well with each other, they cannot both strongly predict success in college.[12]

As discussed above, the existing literature has expressed two conflicting views of convergent validity across performance metrics. On the one hand, any combination of high validity of measurement of similar definitions of "social responsibility," social forces that encourage imitation, herding by raters due to career concerns, or companies that vary in their effectiveness of greenwashing leads to these hypotheses:

---

[12] Recall that convergent validity is not the same as validity for the reasons outlined above. Nevertheless, the relationship can be close. To see this, consider the case where the two measures have the same *validity* (that is, correlation with the true construct), and independent measurement errors with identical variances. In that case, the correlation between the two metrics (their *convergent validity*) equals their validity.

H1: SRI raters will have high convergent validity prior to adjusting for explicit

differences in methods and goals.

On the other hand, any combination of the desire for product differentiation or largely

uncorrelated measurement error leads to:

H1′: SRI raters will have low convergent validity prior to adjusting for explicit

differences in methods and goals.

However, if the diversity in ratings were due to the raters' desire to measure different constructs,

then we have:

H2: Adjusting for explicit differences in methods and goals should greatly increase

measured convergent validity compared to not adjusting for explicit differences in

methods and goals.

At the same time, if significant measurement error exists or if the raters are influenced by the

desire to be distinctive for its own sake or due to measurement error that is not highly correlated,

we have:

H2′: Measures of convergent validity will remain low even after *adjusting* for explicit

differences in methods and goals.

Predictive validity

Even if we find high convergent validity among leading social ratings, they might all have common measurement error. Thus, it is important to also assess the predictive validity of social ratings as well. In other words, are the ratings accurate in assessing which companies will be responsible and irresponsible in the future? While it is difficult to know exactly what investors want when they shop for a "socially responsible" company, it is safe to assume few socially minded investors do want to invest in enterprises that commit massive fraud against investors (e.g., Enron and Tyco), illegally exploit consumers (e.g., Enron during California's electricity crisis), kill thousands of nearby residents (Union Carbide in Bhopal), destroy a local ecosystem (Exxon Valdez), or discriminate massively against female employees (State Farm's anti-discrimination settlement of over $100 million).

Thus, the ability of the social metrics to predict major scandals in the near future is an additional criterion for a good rating (*predictive validity*). Regardless of the motives of investors (described in Appendix 2), all desire predictive validity for social ratings (for a full explanation, see Chatterji, Levine, and Toffel, 2007).

One recent analysis has claimed that corporate scandals and the resulting erosion of trust in top management is a major driver of the growth of SRI. "Many investors are attracted to an investment process based on research that goes deeper and considers qualitative information designed to identify corporate character."[13] A consistent corporate ethical character can arise from several channels: some founders may use their discretion to establish norms of high (or

---

[13] Steven J. Schueth, "SRI in the US", http://www.firstaffirmative.com/news/sriArticle.html. Similar claims are made at
(http://www.novonordisk.com/sustainability/socially_responsible_investment/socially_responsible_investment.asp)

low) social responsibility; some corporate strategies may be based on high (or low) social responsibility, particularly in niche markets; companies with high visibility or responsiveness to stakeholder sentiment may find it optimal to be consistently high on many dimensions of social responsibility. Any of these stories of a consistent "corporate character" coupled with the postulate of valid SRI lead to the hypothesis:

> H3: SRI ratings have high predictive validity, with fewer scandals at firms with high
>
> social ratings than at otherwise similar firms with low social ratings.

Yet critics of SRI rating firms have noted, "Many of the recently disgraced companies, including Anderson, Enron, WorldCom, Adelphia, Tyco, and Tenet Healthcare were favorites of social funds." (Entine 2003: 357) These critics posit that social ratings are largely noise, with little relationship to true corporate social behavior. Our goal in this section is to test whether this anecdotal critique generalizes; Entine's (2003) analysis and anecdotes suggest:

> H3′: SRI ratings have low predictive validity.

Data

We used data from social raters, the Gompers, Ishii, and Metrick (2003) / IRRC index of weak governance, and a list of scandals combining data from the *Corporate Crime Reporter, The Wall Street Journal, The United States Government Accountability Office(GAO),Corporate Environmental Profiles Database(CEPD)* and our own research. We present a detailed view of our data and summary statistics in Table 1.

Social ratings

We used data from several of the major raters: KLD, Calvert, FTSE4Good, DJSI, and Innovest.

For each rater, we used their list of "approved" stocks drawn from a universe most similar to the

Russell 1000. These included the KLD Large Cap Social Index, the Calvert Social Index, the

FTSE4Good Index, the DJSI World Index, and Innovest's 17 U.S.-based firms in its "Top 100

Leaders in Sustainability".

In addition to membership, we had more detailed data for all firms rated by KLD and some firms

rated by Calvert and DJSI. For KLD, we had 60 detailed subscores rating the social performance

of each company. Each detailed subscore is a 1/0 indicator for a strength or concern on topics

such as recycling and emissions. (For a complete list of subscores, see www.kld.com.) For

Calvert, we had five high-level subscores for only the 100 largest firms they rate. Calvert gives a

1 to 5 score for its five top-level categories: the Environment, Workplace, Business Practices,

Human Rights, and Community Relations. For DJSI, we had within-industry rankings for the

top 10% of firms in each industry plus one "runner-up" per industry. Appendix 1 provides

additional description of the raters and the data sources.

IRRC index of weak governance

In our predictive validity analysis we supplemented our social ratings with an index of weak

governance that Gompers Ishii, and Metrick (2003) created using data from the Investor

Responsibility Research Center (IRRC). [14] The IRRC was founded in 1972 and provides

information to more than 500 institutional investors on corporate governance and social

---

[14] http://finance.wharton.upenn.edu/~metrick/governance.xls

responsibility.[15] The index ranges from 0 to 24, with higher numbers being associated with more provisions entrenching managers and, thus, poorer governance. The IRRC (Gompers, Ishii, and Metrick, 2003) provides this rating for 1990, 1993, 1995, 1998, and 2000

The governance score is different than the other ratings discussed above because it focuses directly on corporate governance and is much more transparent in its methodology. By comparing the KLD scores and IRRC governance scores across firms involved in major scandals, we gained insight into whether social ratings have the ability to separate responsible companies from irresponsible ones.

Scandals

We first identified firms that underwent "major" scandals using 5 sources. Our first data source was the Corporate Crime Reporter's (http://www.corporatecrimereporter.com/) list of the 100 largest corporate crimes/scandals of 1990-1999. Our second source was the Corporate Environmental Profiles Database[16] (CEPD) list of the 100 largest environmental (both oil and chemical) spills and accidents from 1991 to 2003. To identify more recent scandals, we next used Internet searches and the Equal Employment and Opportunity Commission (EEOC) annual reports on the lawsuits it filed. Our fourth source of data came from the General Accounting Office (GAO): it reported on earnings restatements between 1997-2003.[17] The database includes restatements data from Lexis-Nexis searches, and focuses on restatements associated with "accounting irregularities". Finally, we used the Wall Street Journal's "Perfect Payday Option

---

[15] The Investor Responsibility Research Center Webpage, (http://irrc.com/index.html) Last accessed November 15th. 2006

[16] http://www.irrc.org/prod_serv/products_environmental2.htm

[17] http://www.gao.gov/new.items/d03395r.pdf

Scorecard"[18] to identify firms who were suspected of manipulating the grants of their stock options.   In total, we identified 1355 plausible scandals: 100 from Corporate Crime Reporter, 100 from CEPD, 71 from our web searches, 919 from GAO, and 165 from the Wall Street Journal's "Option Scorecard." In our methods section, we explain the matching process to KLD ratings and governance scores, which significantly reduced the numbers of scandals used in our analysis. At the end of this process, we identified 218 firms that had scandals during our sample period and also had earlier social ratings we could match.

Other than the GAO restatements data, these methods disproportionately identify very large scandals as measured by the size of the fine and the publicity surrounding the event, which biases our list towards large companies. To account for this feature of the data, in the analysis below we carefully controlled for company size.

Our measures of scandals have several other limitations. We missed some scandals when unethical behavior was legal. For example, consider employment discrimination in the 1950s or massive campaign contributions to political parties in recent presidential elections when the donating company intends to corrupt the political process. Similarly, we missed companies that simply had not (or not yet) been caught. In the other direction, we falsely classified some firms as having had scandals when their behavior was not socially irresponsible. For example, not all restatements or accusations of criminal conduct indicate a lack of social responsibility. However, we do not know why such failures to be caught or such innocent restatements of earnings should be correlated with our measures of social responsibility or weak governance. Thus, the presence

---

[18] http://online.wsj.com/public/resources/documents/info-optionsscore06-full.html (Last accessed September 2007)

of these types of measurement error reduces the statistical power of our tests, but need not create bias.

Methods

We first discuss how we measure convergent validity (that is, how ratings resemble each other) and then predictive validity (how well ratings predict future scandals).

Convergent validity
Measuring convergent validity was difficult because there is no natural definition of "high" or "low" agreement between two ratings. This is a standard problem in examining correlations and validity metrics in other spheres. We used several objective, yet imperfect, benchmarks to describe convergent validity as "high" or "low." Each reader will need to decide if the level of convergent validity is high or low in substantive terms.

A related problem is that there are numerous measures of similarity among the discrete and continuous measures we analyze. Each of these measures, in turn, has problems in capturing the substantive meaning of convergent validity. For example, we examined the extent to which memberships in the several SRI indices overlap. A measure of the share of overlapping membership can be misleading if one index covers only a few firms. For example, if one index includes 500 firms from a universe of 1000 and 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 top half of the first index. Substantively, it would indicate that the second index's top 1% (roughly 3 standard deviation outliers) did not all fall in the top half of the first index. Thus, while we present the

raw figures on overlap, we emphasize measures that are invariant to the number of members in each index

We also performed several tests of the statistical significance of the overlap of memberships and correlation across measures. Statistical significance can be misleadingly encouraging about the economic importance of a relationship, in that it tests a null hypothesis of *zero* relation between two measures of social responsibility. Convergent validity requires a *strong* relation, not just one different from zero.

Membership and Tetrachoric Correlations

We first examine the overlap of the several indices. As noted above, at any level of convergent validity, overlap tends to increase with the size of the indices. In addition, "overlap" is not in units familiar to most social scientists. We can get a better feel for the quantitative magnitude by switching to correlations adjusted for the dichotomous nature of the data. We begin by assuming a standard measurement model:

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

where:

       R_{ij} is the unobserved continuous rating measured by an SRI firm *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."

We assume the true level of social responsibility of firms is also distributed normally. We also assume that e_{ij} is normally distributed and that errors are independent across raters and firms – we discuss correlated measurement error below. We assume that measurement error of different raters has identical variance, which we normalize to unity. Without loss of generality we normalize the mean true responsibility level $T_i = 0$.

For most of our raters, we observe only the discrete measure of whether SRI rater j has firm i in its membership: M_{ij}. We assume that the discrete membership rating M_{ij} equals one when the unobserved continuous rating R_{ij} is above SRI rating agency j's cutoff for membership ($Cutoff_j$):

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

Variation in $Cutoff_j$ is driven by each rating agency's desired size for its membership (for indices of fixed size) or by a rater's view of the acceptable minimum (for raters with an absolute bar). We used maximum likelihood techniques to estimate the correlation of two raters' unobserved continuous ratings (that is, the squared coefficient, b). These estimated correlations are known as tetrachoric correlations.

In addition to membership, we also analyzed the social subscore data from KLD and Calvert. To do so we combined KLD's 63 domain-specific sub-scores into a KLD Score predicting KLD LCS membership and use Calvert's 5 scores to create a Calvert Score for predicted membership in the Calvert Social Index. We estimated these scores as the predicted probability of

membership of firm *f* for each social rater *r* depending on all the subscores of that firm over the domains $i = 1$ to 5 (for r = Calvert) or 63 (for r = KLD):

2)   Membership in relevant $\text{Index}_{rf} = F (\Sigma_i \beta_i \text{ sub-score}_{rfi})$

Here F (.) is the logit function and the β_i variables capture the importance of each sub-score on predicting membership in KLD's LCS or the Calvert Index. The KLD or Calvert Score is a weighted average of the KLD or Calvert sub-scores transformed to range from zero to unity. The weights were chosen to best predict membership in the relevant index, and the units of each score are the predicted probability of index membership. Companies that are excluded by KLD screens, like cigarette makers, were assigned a zero score. Results from sample regressions are in appendix 3.

Adjusting for explicit differences in methods
One measurement challenge is adjusting for the explicit differences between raters. For example, KLD differs from the other ratings in the screens it uses and in its lack of norming by industry. Our adjustments modified KLD's underlying data to mimic the explicit screens and industry norms used by the other funds.

The "Calvert style" and "FTSE style" scores score use a similar regression to equation (1) but set the score equal to zero only for the screens used by Calvert or FTSE. The "DJSI/Innovest style" score uses the probabilities from equation 1 without any adjustments for screens (as DJSI and Innovest do use explicit screens).

Furthermore, we included industry dummies when estimating equation 2 for these 3 scores, because the other indices norm their ratings by industry. By adding in industry dummies, we are also norming KLD's Score by industry. The end result provides us with 4 scores that are adjusted for differences across raters and can be used for meaningful comparisons of convergent validity holding constant explicit differences in rating strategy.

Predictive Validity

In our context, predictive validity means that firms with high social scores have lower probabilities of being in scandals, holding all else constant. We identified 218 scandals in the universe of firms rated by KLD at least three years prior to the scandal. 92 of these scandals came from the Wall Street Journal, 72 from the GAO, 14 from the Top 100 Corporate Criminals, 13 from CEPD's list of largest spills, and 27 from our own Internet searches. For each scandal firm we identified a comparison firm with similar employment three years previous and the same 2-digit industry that was also in KLD's universe. In those cases where we could not find data for the firm 3 years prior to scandal, we used a 2 year lag. This method added 18 firms to our total list of scandals and all results are robust to the exclusion of these 18 firms. We then measured membership in the Domini 400 three years prior to the scandal for the scandal firm and the comparison firm. We repeated this procedure for firms in the universe of Gompers, Ishii, and Metrick ratings. Our tests of predictive validity involved asking whether scandal firms were systematically less likely to be in KLD's Domini 400 three years prior to the scandal and had systematically higher indices of weak governance than the control firms (again, roughly 3 years prior to the scandal).

In the KLD analysis, there were 218 scandal firms and 210 controls. We used the Domini 400 because it is more selective index than the LCS and because data is available beginning in 1991, allowing us to analyze more scandals. For the Gompers, Ishii, and Metrick(2003) / IRRC governance score analysis, we had 84 scandal firms and 80 controls. Once again, we used data 3 years prior to scandal, and if unavailable, data 2 years prior to scandal.

Results

Convergent Validity

We first discuss overlap among the memberships of the various SRI indices and correlations among scores and sub-scores. In the next section we examine correlations among the several continuous measures we construct as well as the relations between discrete measures of index membership. We conclude this sub-section by examining the extent to which correlations are low due to purposeful differences in index measurement that we are able to adjust for versus other sources of divergence (either purposeful or erroneous).

Overlaps of membership and Tetrachoric Correlations

First, we describe how well the memberships of the 5 different indices correlate. In Table 2, we report the tetrachoric correlations implied by the overlap in memberships. The overlap in membership underlying these tetrachoric correlations are in Appendix 4 (for example, the share of KLD's LCS members that are in Calvert's index). Recall that the tetrachoric correlations rise with overlapping membership (from which they are derived), although the correlations are invariant to changes in index size. Because the results can be difficult to interpret, we first briefly summarize the results and then explain them at length.

Summary of the results

Nine of the 10 tetrachoric correlations were positive (that is, all but an insignificantly negative correlation between the membership of KLD's LCS and Innovest) and four of the 9 positive

correlations were statistically significant. The average of the ten distinct tetrachoric correlations was 0.23, which is just above the median (between 0.13 and 0.22).

To understand what these correlation means in terms of overlap among the indices, consider the overlap of KLD's LCS and FTSE4Good (which have the median tetrachoric correlation of 0.22), where all overlap calculations rely solely on firms in the universes rated by both raters. Out of FTSE4Good's top 78 firms, 79% are in KLD's LCS as are 65% of the 513 FTSE nonmembers. The higher KLD membership share among FTSE4Good's members than its nonmembers is statistically significant. At the same time, it is not impressive that when FTSE chooses only an elite 12% of firms, one fifth of that elite are not in KLD's top two thirds.

To think of the mean correlation of 0.23 another way, it implies if one rater measured a company as two standard deviations high or low on social responsibility (that is, in the top 5%), the typical other rater rated that same company less than a half a standard deviation high or low.

Thus, there is not strong evidence that all the raters are measuring a single construct with high accuracy. These results support H1' in that the ratings exhibit low convergent validity – at least prior to any adjustments for differences in emphasis.

Detailed Results
KLD's Large Cap Social Index membership correlated fairly well with Calvert, with an implied tetrachoric correlation of the underlying ratings of 0.69. (Correlations are statistically significant at the 5% level unless marked as not significant.) To see the source of this strong correlation, note that 89% of the 493 Calvert members, but only 47% of the 490 Calvert nonmembers, are in the LCS.

The KLD-Calvert relation was by far the strongest of the 10 correlations. The tetrachoric

correlation for KLD's LCS was only 0.01(not significant) with DJSI, 0.22 with FTSE4Good, and

-0.22 (not significant) with Innovest. These weak correlations arose because 65% of DJSI's 80

members and waitlist companies are in LCS, as opposed to 60% of the DJSI "others;" 79% of the

FTSE4Good top 78 are in LCS, and 65% of the 513 FTSE nonmembers; and 47% of the 17

Innovest members and an even greater 67% of the 542 non-members are in LCS. Calvert was

almost uncorrelated with the other three ratings, with correlations of only 0.07 (not significant)

with DJSI, 0.13 (n.s.) with FTSE4Good, and 0.09 (n.s.) with Innovest.   In contrast, DJSI had

large and statistically significant tetrachoric correlations with both FTSE4Good (0.53) and with

Innovest (0.54).[19] Finally, FTSE4Good and Innovest had a modest 0.23 correlation (n.s.).

Are differences in rankings due to explicit differences in method?
In this section we examine whether explicit differences in method, actual use of screens, and

implicit differences in weights on sub-scores account for divergences in membership and ratings

across the several indices. This calculation is most straightforward if we examine the relation

between KLD Scores and the other indices, as KLD has the most distinctive set of criteria. First,

KLD is the only index that rates firms across industries; the others all claim to be set relative to

industry benchmarks. Second, KLD has more explicit screens than the other funds.

If we examine Calvert, for example, the 493 members of Calvert's index have a mean KLD score

that is 0.30 higher than the 400 nonmembers. If we use Calvert's screens instead of KLD's and

---

[19] To summarize these results, 9 of the 10 correlations are positive (that is, all but the insignificantly negative
correlation between KLD's LCS and Innovest), and 4 of the 9 positive correlations are statistically significant. The
average of the ten distinct tetrachoric correlations is 0.23. That correlation implies if one rater measured a company
as two standard deviations high or low on social responsibility (that is, in the top 5%), the typical other rater rated
that same company less than a half a standard deviation high or low.   Thus, there is not strong evidence that all the
raters are measuring a single construct with high accuracy.

we norm by industry (as Calvert does, but not KLD), the gap declines by a small and not statistically significant amount to 0.29.

The unimportance of adjusting for different (or no) screens and for industry norms repeats for the other indices. FTSE4Good members have a KLD Score that is a tiny 0.05 larger than for non-members. This gap rises by a small and not statistically significant amount to 0.08 when we switch to using FTSE4Good's screens (not KLD's) and industry norming. DJSI and Innovest members also have slightly higher KLD Scores than non-members (0.02 and 0.15). These gaps shift slightly when we drop screens and norm by industry, rising slightly to 0.05 for DJSI and declining slightly 0.03 for Innovest. No changes are economically or statistically significant. (We repeat this exercise for the several continuous social ratings in Appendix 5.)

In short, there is no evidence that differences among ratings are due to explicit the differences in measurement strategy that we can adjust for. These results support H2′: Convergent validity remains low even after adjusting for explicit differences among the raters.

Predictive Validity
Now that we have considered convergent validity, we evaluate the predictive validity of the ratings. Specifically, we test whether KLD's Domini 400 membership and ratings and the Gompers, Ishii, and Metrick (2003)/ IRRC index of weak governance in one year predicts scandals three years later.

Recall that we first matched scandal firms by 2-digit SIC code and by employment 3 years prior to public information on the scandal. In those cases where we could not find data 3 years prior to

the scandal, we used data 2 years prior to the scandal. Our matching on size was successful; the

mean log of employees for both scandal and comparison firms was 9.06.

Does Domini 400 membership or KLD sub-scores predict fewer scandals?
If membership in KLD's Domini 400 strongly predicts fewer scandals, we expect the pool of 218

scandal firms to have fewer Domini members than the pool of 210 matched non-scandal firms.

In fact, 35% of the scandal firms and 36% of the control firms are in the Domini 400 and the

difference is not significant, implying that the predictive validity of Domini 400 membership is

weak. (Table3A) At the same time, the confidence interval on the odds ratio is very wide so our

test does not have power to rule out economically meaningful effects -- in either direction.

When we examined KLD sub-scores we found a shred of evidence that KLD has predictive

power, but it was also coupled with a shred of evidence that it does not. Specifically, we

summed the number of concerns in the domains of Community, Diversity, Employee Relations,

Environment, and Product. The mean scandal firm had 1.4 of these five concerns, while the

mean control firm had only 1 concern. The difference is statistically significant at the 5% level (t

= -2.96, matched t test; results for KLD sub-scores and predictive validity are in Appendix 6).

While it is encouraging for KLD sub-scores to have some predictive validity, this correlation is

not convincing for two reasons. First, we also examined whether companies with more KLD

strengths had fewer scandals. Of the five possible strengths, the average scandal firm had 1.7

while the average non-scandal control had 1.3 -- that is, scandal firms averaged nearly half a

strength *more* than did controls without a scandal. The difference is significant (t = -2.17), and

the magnitude of this gap in strengths (that is inconsistent with the theory that socially

responsible firms have fewer scandals) is as large as for the gap in concerns (that was consistent with that theory). Finally, when we use all KLD sub-scores or just the KLD Score (the index of KLD sub-scores that best predicts KLD membership, estimated in Appendix 3) to predict scandals, there is not statistically significant relationship indicating predictive validity.

Does the Gompers, Ishii, and Metrick / IRRC index of weak governance predict fewer scandals?

Our tests of predictive validity for the Gompers, Ishii, and Metrick index of weak governance followed the same method as our test using Domini 400 membership. If the index of weak governance strongly predicts fewer scandals, we expect the scandal firms to have a higher mean index (that is, worse governance) than the non-scandal firms. We found that the scandal firms had a mean index of weak governance of 9.74. (Table3B). For the matched comparison firms that had no scandal, the mean index of weak governance was 10.06. The weaker governance of the control firms is not statistically significant (t = 0.8024) and is in the opposite direction predicted by theory. However, while we found no evidence that the index of weak governance predicts scandals, we note that our statistical power is modest, in that the 95% confidence interval ranges from scandal firms having roughly ½ point stronger governance to 1½ points weaker governance than comparison firms.

Overall, these results support H3′ (that social ratings have low predictive validity) rather than H3. At the same time, the results do not support the generalizability of John Entine's anecdotal evidence (quoted above) that firms with high social scores are *more* likely to have scandals.

Conclusion

Summary

For convergent validity we found evidence that these ratings firms are rating related constructs in the sense that the correlations in underlying ratings on average correlate around 0.23. This figure is below unity due to purposeful differences and measurement error. If one of the raters measured the underlying construct of social responsibility with more precision than the others, we would observe consistently higher levels of correlation for that rater. Our current results indicated no such pattern. In our own view of what correlation corresponds to "high" convergent validity, we found support for Hypothesis 1', but for each reader it will depend on the interpretation of a 0.23 correlation between ratings.

Our differences in measurement remained even after adjusting for explicit distinctions in measurement. Thus, we also found support for Hypothesis 2' because convergent validity remained low to modest after our adjustments for explicit differences in rater goals and methods. Thus, we were left with the strong suspicion that measurement error accounted for a significant share of the variance in raters' true ratings of corporations' social performance. However, future research will be necessary to estimate the exact amount of measurement error in these social ratings.

Our results on predictive validity were even less encouraging than our results on convergent validity. Our results provided support for Hypothesis 3'' that Domini 400 membership and the Gompers, Ishii, and Metrick index of weak corporate governance have weak predictive validity. Neither a narrow focus on governance or a broad measure of social responsibility (including

charitable giving, environment impact, product safety, etc.) seems to distinguish firms that will have major scandals from those who will not.

While we found little support for the theoretical predictions of strong convergence through imitation, we can yet accept theories that predict divergence stemming from competitive reasons or the quest to be distinctive. After all, we found strong evidence that significant measurement error likely exists, and future research can advance our understanding of how much impact this error has on the social ratings in this study.

Discussion

Our findings are consistent with several interpretations, each with implications for SRI raters and for investors. To the extent that the low convergent validity of social ratings reflects different criteria that investors understand and respond to, diversity is a desirable outcome of market forces and should persist. For example, we expect both funds that avoid contraception providers (perhaps on religious grounds) and funds that that search out contraception providers (perhaps with the goal of empowering women) to meet the goals of specific niches of investors. At the same time, accounting for explicit differences across raters did not consistently improve the fit of the ratings, making us doubt that most of the divergence in scores is due to purposeful differences in approach that investors understand.

Our results are also relevant for other areas of finance where ratings are crucial, whether in the recommendations of sell-side analysts or the ratings of corporate bonds. Our study is among the first to test for convergent validity among social ratings, and it would be interesting to examine the parallels between the early stages of the financial industry and the socially responsible

investing industry. Our sense is that the current state of SRI is similar to other ratings systems in their early stages of development.

There are several important possible extensions to this research. First, we could include more ratings from additional raters in the socially responsible investing industry. In addition, along with considering a particular rater's screens, we could compute the ratings using our own definition of screens, though we doubt that this would improve the fit or the correlations. It would also be interesting to identify subscores from more raters; we had access to only those of KLD for this research project. Furthermore, we could also address the construct validity issue by unpacking the governance scores (e.g. exploiting variation by state) and rater scores, and looking at the measurement properties of the original surveys.

In terms of predictive validity, we could add different types of scandals, including the recent option back-dating controversies. With enough scandals, we could potentially separately test for environmental subscores as predictors of environmental outcomes and governance subscores as predictors of accounting and governance scandals. The broader results on predictive validity could inform the results of related studies (Chatterji, Levine, and Toffel, 2007), which assessed predictive validity for KLD's environmental subscores in predicting environmental outcomes such as emissions and fines. In the future, we can also examine predictive validity of other raters as they accumulate more historical data.[20]

---

[20] It is also possible that different raters could be measuring different aspects of the same firm's record, leading to low predictive validity if firms are themselves diverse with respect to "responsibility". For example, Exxon Mobil is widely viewed as socially responsible on safety, pollution control and financial reporting, but criticized for their lack of initiative on global climate change issues. (We thank David Vogel for this point)

Our results are consistent with the hypothesis that much of the current diversity in social ratings reflects inconsistent definitions and measures of social responsibility coupled with measurement error – not marketing to distinct niches. As such, our results are broadly supportive of neo-institutionalist theories that when "objective" quality is hard to measure, then differentiation can arise as a form of fashion.

Divergences that reflect experimentation and learning can be a strength of a new industry; the key is to build in learning and a continued stream of validation studies that inform social screening funds and socially conscious investors about the validity of various metrics. It is important that convergent validity rise over time due to increased knowledge, not due to herding or neo-institutionalist forces that promote conformity to arbitrary norms. Finally, if there is not a steady stream of additional research validating the numerous measures (see Chatterji, Levine, and Toffel, 2007 for an example), diversity of measures may continue as both old and new funds market new approaches to measurement.

References

Bagnoli, M., and S.G. Watts

2003　"Selling to Socially Responsible Consumers: Competition and the Private Provision of Public Goods," Journal of Economics and Management Strategy, 12(3), 419–445.

Banerjee, A.

1992　"A Simple Model of Herd Behavior." Quarterly Journal of Economics, Volume 107, Issue 3: 797-818.

Baron, D.P.

2001   "Private Politics, Corporate Social Responsibility, and Integrated Strategy," Journal of
        Economics and Management Strategy, 10(1), 7–45.

Bartels, L.

1988   Presidential Primaries and the Dynamics of Public Choice. Princeton, N.J.: Princeton
        Univ.   Press.

Baum, J.A.C., and H.A. Haveman

1997   "Love Thy Neighbor? Differentiation and Agglomeration in the Manhattan Hotel
        Industry, 1898-1990." Administrative Science Quarterly, Vol. 42, No. 2. pp 304-338.

Basu, K. and Z. Tzannatos

2003   The Global Child Labor Problem: What Do We Know and What Can We Do? **17:** 147-
        173.

Beltratti, A.

2003   "Socially Responsible Investment in General Equilibrium." SSRN working paper,
        http://papers.ssrn.com/sol3/papers.cfm?abstract_id=467240

Bikhchandani, S., D. Hirshleifer, and I. Welch

1992   "A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades."
        Journal of Political Economy, Vol. 100: 992-1026.

Bourdieu, P.

1984   A Social Critique of the Judgement of Taste. London: Routledge &. Kegan Paul.

Burns, J.L.

2000   "Hitting the Wall: Nike and International Labor Practices." Harvard Business School
        Case Study 9-700-047.

Chatterji, A. and D. Levine

2006   "Breaking Down the Wall of Codes: Evaluating Non-Financial Performance
        Measurement." California Management Review 48(2): 29-51.

Chatterji, A., D. Levine, and M. Toffel

2007   "How Well Do Social Ratings Actually Measure Corporate Social Responsibility?"
        SSRN working paper, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=993094

Crane, D.

1999   "Diffusion Models and Fashion: A Reassessment." Annals of the American Academy of
       Political and Social Science, Vol. 566, The Social Diffusion of Ideas and Things. Pgs. 13-
       24.

Corporate Crime Reporter: List of the largest corporate crimes/scandals of 1990-1999.
(www.corporatecrimereporter.com) Last Accessed November 14th, 2006

DiMaggio, P.J., and W. Powell

1983   "The iron cage revisited" institutional isomorphism and collective rationality in
       organizational fields." American Sociological Review, 48, 147-60.

Entine, J.

2003   "The Myth of Social Investing: A Critique of its Practice and Consequences for
       Corporate Social   Performance Research." Organization and Environment, Vol. 16,
       No. 3: 352-368 http://oae.sagepub.com/cgi/rapidpdf/16/3/352.pdf

Gompers, P.A., J.L. Ishii, and A. Metrick

2003   "Corporate Governance and Equity Prices." Quarterly Journal of Economics, Volume
       118, No. 1: 107-155 http://finance.wharton.upenn.edu/~metrick/gov.pdf

Hawken, P.

2004   Socially Responsible Investing. The Natural Capital Institute.

Hong, H., J. Kubik, and A. Solomon

2000   "Security Analysts' Career Concerns and the Herding of Earnings Forecasts." Rand
       Journal of   Economics, Vol. 31: 121-144.

Kinder, P.

2005   "A Note from Peter Kinder and the KLD Press Release." GreenMoneyJournal.com.
       (Last Accessed September 14, 2005)

Lev, B., C. Petrovits, and S. Radhakrishnan

2006   "Is Doing Good Good for You? Yes, Charitable Contributions Enhance Revenue
       Growth," Working Paper, New York University Stern School of Business.

Lyon, T. P., and J. W. Maxwell

2005   "Greenwash: Corporate Environmental Disclosure under the Threat of Audit." Working
        Paper, University of Michigan.

March, J. G. and J.P. Olsen

1976   Ambiguity and choice in organizations. Bergen: Universitetsforlaget.

Maxwell, J.W., T.P. Lyon, and S.C. Hackett

2000   "Self Regulation and Social Welfare: The Political Economy of Corporate
        Environmentalism,"   Journal of Law and Economics, 43 (2), 583–618.

Merton, R.

1987   "A Simple Model of Capital Market Equilibrium with Incomplete Information." Journal
        of Finance, Vol. XLII, No. 3.

Meyer, Marshall W., and V. Gupta

1994   "The Performance Paradox," in Research in Organizational Behavior, L. L. Cummings
        and B. M. Staw (eds.), Greenwich, CT, JAI Press, Volume 16: 309-369

Orlitsky M., F.L. Schmidt, and S.L. Rynes

2003   Corporate social and financial performance: a meta-analysis. Organization Studies, 24
        (3):403.

Rao, H., H. Greve and G. Davis.

2001   "Fool's Gold: Social Proof in the Initiation and Abandonment of Coverage by Wall Street
         Analysts." Administrative Science Quarterly. 46: 502-526.

Rosen, B.N., D.M. Sandler, and D. Shani

1991   "Social Issues and Socially Responsible Investment Behavior: A Preliminary Empirical
        Investigation," Journal of Consumer Affairs, 25 (2), 221–234.

Scharfstein, D.S., and Stein, J.C.

1990   "Herd Behavior and Investment." American Economic Review 80, pg. 465-479

Scott, W.R.

1995 Institutions and Organizations. Thousand Oaks, CA; Sage

Sharfman, M.

1996   "The Construct Validity of the Kinder, Lydenberg & Domini Social Performance Ratings
        Data." Journal of Business Ethics 15: 287-296

Social Investment Forum: *2003* Report of Socially Responsible Investing Trends in the United
States, 2003.

Stata, "Tetrachoric Correlations for Binary Variables," in Stata Reference R-Z, 2006, Plano TX:
426-434.

Staw, B.M. and L.D. Epstein

2000   "What Bandwagons Bring: Effects of Popular Management Techniques on Corporate
        Performance, Reputation, and CEO Pay." Administrative Science Quarterly, 45: 523-556

The Financial Statement Restatement Database, United States General Accounting Office, 2003
[Annex 2 of http://www.gao.gov/new.items/d03395r.pdf].

Vaaler, P.M. and G. McNamara

2004   "Crisis and Competition in Expert Organizational Decision Making: Credit-Raters and
        Their   Response to Turbulence in Emerging Economies." Organization Science, Vol. 15,
        No. 6. pp. 687-703.

Waddock, S.

2003   "Myths and Realities of Social Investing." Organization and Environment, Volume 16,
        Number 3: 369-380.

Zitzewitz, E.

2001   "Opinion-producing agents: career concerns and exaggeration." Stanford Business
        School Working Paper

Table 1: Summary Statistics

Membership in Social Indices (2003-2005)	In	Out	Universe (N)	Universe name
KLD LCS membership	670	313	983	Russell 1000
Calvert Membership	493	490	983	Russell 1000
FTSE4Good Membership	77	501	578	FTSE All World USA
DJSI Membership (including waitlist)	77	834	911	Dow Jones World Index
Innovest Membership-(Top 17)	17	485	502	"Innovest Universe"

Continuous Measures of Responsibility	Mean	SD	Min	Max	N
KLD Score = Predicted probability of KLD LCS membership; from Appendix 1					
Using 63 subscores (0 or 1) and KLD screens	0.67	0.35	0	1	1000
With common KLD and Calvert screens	0.65	0.34	0	1	1000
With common KLD and FTSE screens	0.4	0.29	0	1	1000
With common KLD and DJSI/Innovest Screens	0.45	0.26	0	1	1000
Calvert Score = predicted probability of Calvert Index membership based on 5 subscores (each 1 to 5), from Appendix 1. (Calvert subscores are available only for the 100 largest firms)	0.53	0.395	0.0000863	0.9997	100
DJSI rank within 22 industries. Ranks are available for DJSI's top 10% of each industry (so range = 90-100 percentile) and 22 waitlisted firms	0.93	0.029	0.895	0.9857	88
G Index of Weak Governance (# of 24 items Gompers et al. coded as weak corporate governance)	9.1	2.77	2	19	704

Scandal measures 1993 to 2006	# Matched to KLD and Compustat
Earnings Restatements 1993-2006	72
Largest 100 spills 1993-2006	13
100 Corporate Criminals	14
Media Search	27
WSJ Perfect Payday	92
Total # Scandals used in analysis	218

^ Obtained through logit post estimation, where subscores the perfectly predict success and failure are multiplied by 1.5 times the largest coefficient (positive or negative) and subtracted out

^^ Same as above, except without SP 500 as a control in the logit equation

^^^ Same as above, except with explicit screens, like alchol, tobacco, etc., included in the logit equation and explictly screened out companies set to zero

^^^^ Same as above, except with industry controls in the logit equation

Table 2: How much do the memberships of top Social Ratings funds intersect?

Tetrachoric correlations

Index and [maximum N in index]	Calvert	DJSI	FTSE4Good Top 100	Innovest Elite 18	Mean correlation of this index with the other 4
KLD's LCS	TC=0.6856 *; Chi2=198.83 *, LO=9.12 N=983;	TC=.01; Chi2=0; LO=1.02; N=911;	TC=0.2196 *; Chi2=5.89 *; LO=2.03; N=578;	TC=-.22; Chi2=2.93; LO=.438; N=560;	0.1738
Calvert Social Index [607]	X	TC=0.07; Chi2=0.77; LO=1.24; N=911;	TC=0.13; Chi2=2.64*; LO=1.50; N=578;	TC=0.09; Chi2=0.41; LO=1.38; N=507;	0.2439
DJSI top 10% plus 1 in each industry [88]	X	X	TC=0.53 *; Chi2=45.53*; LO=5.98; N=570;	TC=0.54 *; Chi2=26.06 *; LO=8.52; N=534;	0.2875
FTSE4Good Top 100 [101]	X	X	X	TC=0.23; Chi2=2.58; LO=2.44; N=392;	0.2774
Innovest US firms in Innovest's global top 100 [17]	X	X	X	X	0.16

Notes: In each cell, TC means tetrachoric correlation, Chi2 is Chi Squared Statistic, LO is the log odds ratio, which is the log of the ratio of the likelihood of inclusion in an index over exclusion.
and N is the intersection of the universes of the two rating agencies
As described in the text, tetrachoric correlations are similar to standard correlations, but are adjusted for the dichotomous nature of the data.
Membership is from 2005.
* Significant at the 5% level

Table 3: Predictive Validity Analyses

Table 3A: Does membership in the Domini 400 Index predict fewer scandals?

Scandal in Year t	Member	Non-Member	Column Total	N (row)
	Domini 400 membership status in Year t-3			
Scandal firms[a,b]	35%	65%	100%	218
Non-scandal comparison firms	36%	64%	100%	210
N(column)	152	276		428

McNemar chi2=19.98 testing whether column shares are equal (n.s.); Log Odds Ratio=1.87

a: Firms with scandals are all firms rated by KLD that two or three years later had earnings restatements, one of the largest environmental spills, or were members of a list of major scandals collected by the authors.
b: The same control observation can be matched to more than 1 treatment observation

Table 3B: Do IRRC Governance Scores predict fewer scandals?

G Index of weak governance ranges from 1-16, with higher numbers associated with weaker corporate governance

Correlation between G Index of Weak Governance and Scandals $_{t+2}$	-0.0629 N=164
Mean G Index of Weak Governance for Scandal Firms	*9.74 N=84*
	(sd=2.84)
Mean G Index of Weak Governance for Matched Non-Scandal Firms	*10.06* N=80 [c]
in same 2-digit industry and closest # of employees in year t-2	(sd=2.3)
Gap	0.324
	(sd=0.40)
t statistic on gap (n.s.)	0.8024

c: The same control observation can be matched to more than 1 treatment observation

Table 4: Gap between index members and non-members in predicted probability of KLD LCS membership

Gap in KLD Score (= Predicted probability of KLD LCS membership) between index members and non-members
SE in parentheses, maximal gap possible with membership of this size in {braces}.

Screens	Industry norming	KLD LCS	Calvert	FTSE4Good Top 100	DJSI	Innovest
1. KLD style: Screened-out firms get Score = 0. Others get probability predicted based on 63 subscores (0 or 1).		0.55 (0.02) * {0.73}	0.30 (0.02)* {0.52}	0.05 (0.05) {0.40}	0.02 (.04) {0.37}	0.15 (.09) {0.40}
KLD screens get Score = 0	no					
2. Calvert style: Screened-out firms (usingKLD measure of Calvert screens) get Score = 0. Others get probability predicted based on 63 subscores (0 or 1). Logit, but not prediction, also has industry dummies.		0.53 (0.02)* {0.70}	**0.29 (0.02)* {0.52}**	0.02 (0.04) {0.40}	0.02 (0.04) {0.40}	0.11 (.09) {0.42}
Calvert Screens get score = 0	yes					
3. FTSE4Good style: Screened-out firms (using KLD measure of FTSE screens) get Score = 0. Others get probability predicted based on 63 subscores (0 or 1). Logit, but not prediction, also has industry dummies.		0.37 (0.02)* {0.43}	0.24 (0.02)* {0.50}	**0.08 (0.04)* {0.60}**	0.11 (.04)* {0.56}	0.02 (.08) {0.39}
FTSE4Good screens get Score = 0	yes					
4. DJSI and Innovest style: All firms get probability predicted based on 63 subscores (0 or 1). Logit, but not prediction, also has industry dummies.		0.29 (0.02)* {0.58}	0.22 (0.02)* {0.44}	0.02 (0.03) {0.51}	**0.05 (.03) {0.52}**	**0.03 (.07) {0.40}**
No screens	yes					
Change in gap between KLD style and "Actual" (In Bold)			.01 (.03)	-0.03 (.05)	-0.03 (.05)	.12 (.11)
N =overlap of universes of KLD and other index		983 *	983	578	911	502

* The universe for KLD LCS is the Russell 1000, but we have social ratings for 983 of the 1000 firms in the index

Notes:
A firm's Score is the predicted probability a firm would be included in the KLD LCS index. We vary the formula for the Score to approximate the methods used by indices other than KLD.
If the rater in question would screen out the firm, the firm is given a Score of zero.
Screens for the other indices are proxied with KLD's measures of that screen. For example, FTSE4Good has a screen on gambling, but due to data limitations we use KLD's, not FTSE's definition of "screened out for gambling."
If not screened out, the company's score is the predicted probability of membership in the KLD LCS based on its 64 KLD subscores. See appendix for an example of this regression.
Industry norming of KLD score is carried out by including industry dummies in the logit equations predicting KLD LCS membership,
but dropping the dummies when predicting membership.
The {maximal gap possible} for an index of size n in a universe of size N compares the mean score of the n highest Scores
and the (N-n) lowest scores; that is, the gap in Scores that would arise if KLD's subscores were all that were used in choosing the index in question.

Appendix 1: Social Raters

In this appendix we briefly describe the 5 social raters we analyze.

KLD

KLD is one of the oldest (1988) and most influential social raters with $8 billion invested in funds based on its index.[21] KLD's objective is "to provide global research and index products to facilitate the integration of environmental, social and governance factors into the investment process." Many have argued that (e.g., Waddock 1993) KLD creates the highest quality metrics of social responsibility. Several academic researchers have used this data, and scholars generally considered it the standard for measuring corporate social performance.

 KLD assesses seven domains of corporate social performance including: community relations, corporate governance, diversity, employee relations, environment, human rights, and product quality and safety. Within each domain, KLD assigns a positive score (Strength) or a negative score (Concern) based on specific rating criteria. It has explicit screens for alcohol, tobacco, nuclear, firearm, military, and gambling involvement. It reveals no specific weights on its various sub-scores, although we were able to glean some insights into their weights through our analysis. To our knowledge, KLD does not rank firms relative to their industry averages.

KLD researchers study company, government, media and NGO reports to rate over 3000 companies each year. In this study, we used two of KLD's indexes—Domini 400 and the Large Cap Social Index (LCS). To construct the Domini 400, KLD begins with the S&P 500 and eventually chooses the top 250 firms from the index along with 150 other socially responsible

---

[21] Kinder (2005)

firms to comprise its Domini 400 index. We used the Domini 400 index in the predictive validity

analysis because our data begins in 1991 and this we were able to track companies over time. To

construct LCS, KLD begins with the Russell 1000 universe, and selects 662 firms using their

ratings system for inclusion into the index. The LCS was first constructed in 2001, so we only

used this index for the convergent validity analysis but not for our predictive validity analysis

since the Domini 400 index allows for more years of data. We examined both membership in

KLD's indexes and the detailed sub-scores KLD provides for each firm.

Calvert Social Index
Calvert, founded in 1976, manages $12 billion in assets for over 400,000 investors. Calvert

offers over 30 different funds, including the Calvert Social Index. Calvert describes its social

index as a "broad-based, rigorously constructed benchmark for measuring the performance of

US-based socially responsible companies." Calvert selects approximately 600 firms from the

Russell 1000 for inclusion into its social index. These firms are evaluated on the following

criteria: Products, Environment, Workplace, and Integrity. Calvert also reports ratings for the top

100 largest companies, rating the firms on a 1 to 5 scale across 5 categories, the Environment,

Workplace, Business Practices, Human Rights, and Community Relations. The company uses

some of the same explicit screens that KLD does and does not report explicit weights on the

various social criteria. It also rates firms according to average performance in their industry.

Finally, Calvert keeps information on nearly 7000 companies, using Lexis Nexis, trade

publications, government and NGO reports, and various other sources to formulate its decisions.

FTSE4Good
FTSE4Good was launched in 2001 by The Financial Times and the London Stock Exchange, and

donates all of its license revenue ($1.6 million by 2006) to UNICEF. One of FTSE4Good's

primary objectives is "to provide a tool for responsible investors to identify and invest in

companies that meet globally recognized corporate responsibility standards." It judges

companies on the following criteria: environment, stakeholder relationships, human rights,

supply chain management, and "countering bribery". FTSE4Good also includes screens for

tobacco, nuclear, and military concerns. FTSE4Good works with Ethical Investment Research

Services to conduct its research. They collect information from annual reports, company

websites, and other publicly available information.

FTSE4Good divides industries into high, medium, and low impact sectors, and employs different

criteria for each. For each criterion, FTSE4Good evaluates companies on policy, management,

and reporting activities. Depending on the impact of their sector, firms have to meet a fraction of

the recommended criteria to be included in one of FTSE4Good's indexes. In this study, we use

FTSE4Good's US 100, which includes the 100-US based firms in the FTSE4Good index.

Dow Jones Social Index (DJSI)
DJSI was launched in 1999 by Dow Jones Indexes, STOXX Limited, and SAM group and has

sold 56 licenses in 14 nations. In sum, these licensees managed over $4 billon Euro in 2006.[22]

DJSI's goal was to create the "world's first equity benchmark to track the financial performance

of sustainability leaders on a global scale."

DJSI begins with a universe of 2,500 firms from the Dow Jones Global Index and aims to select

the top 10% for the Sustainability World Index. It divides firms into 58 industry sectors, and

places 60% weight on general criteria and 40% on industry-specific criteria. The company places

---

[22] Dow Jones Sustainability Indexes Homepage, (http://www.sustainability-index.com/), Last Accessed October 3rd, 2006)

equal weight on three broad categories, Economic, Environmental, and Social. Importantly, DJSI uses relative rankings by industry, so they are seeking to identify the top 10% in each industry. Thus, it does not use the same screens as other raters, instead trying to identify the best in class, even among "sin" industries like tobacco.

DJSI uses information from 4 sources: company questionnaires, company reports, media reports and interviews with stakeholders, and direct company engagement. DJSI only rates firms that respond to its questionnaire. DJSI works with Sustainable Asset Management (SAM), which employs 20 analysts who spend on average 2 days per company.

The Dow Jones Social Index (DJSI) reported over 80 US-based firms with explicit rankings ratings within their industries. We rescaled the DJSI measures to a percentile within their industry. Recall that DJSI ranks the top 10 percent of each industry plus the highest runner-up. For example, if Dow Jones ranked 3 of 30 firms in an industry we would have relative information on 4 firms (the 3 ranked firms plus the runner up). The other 26 firms would be classified collectively as "not ranked,"

Innovest

Innovest was founded in 1995 and has licensees with $1.1 billion under management in 20 nations. Its web site explains: "At the heart of Innovest's analytical model is the attempt to balance the level of environmentally and socially driven investment risk with the companies' managerial and financial capacity to manage that risk successfully and profitably into the future."[23]

---

[23] Innovest Webpage, (www.innovestgroup.com) Last accessed October 2nd, 2006

Innovest tracks 1750 companies world-wide on 120 individual factors. They focus on 4 areas: EcoValue (environmental issues), Human Capital, Stakeholder Capital, and Sustainable Governance. Innovest analyzes these factors in the context of how they affect financial performance. For the current study, we use the 17 US-based firms in Innovest's Top 100 leaders in sustainability. Innovest claims to have the largest number of analysts in the world, and that over 90% of these analysts hold advanced degrees. Rather than using questionnaires, they interview executives.

Appendix 2: Motives for Investors

While most investors do not use social screens or concerns in choosing their portfolio, a potentially important minority does. Investors have a variety of motives for socially conscious investment, and any particular investor can, of course, hold more than one motive. Investors might choose socially responsible companies because, for example, they associate social responsibility with better financial performance. Prior research has examined how corporate social responsibility (CSR) can have financial benefits for companies by attracting socially responsible consumers (Bagnoli and Watts, 2003), reducing the threat of regulation (Maxwell, Lyon, and Hackett, 2000), improving their reputations with consumers (Lev, Petrovits, and Radhakrishnan 2006), and reducing concern from activists and non-governmental organizations (Baron, 2001; Lyon and Maxwell, 2006).

Alternatively, investors with consequentialist motives hold socially-conscious investments to "invest for their own futures and a better world at the same time" (Entine 2003). It is likely such investors assume that their decision on holding stocks lowers the cost of capital for socially-desirable firms and raises the cost of capital for disfavored firms (in effect "punishing" less responsible firms). For example, marketing materials from socially responsible funds routinely claim that such investments will help improve the world.[24]

---

24 KLD Homepage (http://www.kld.com/about/index.html) Last accessed October 2nd, 2006
 "Those who invest in a socially responsible manner attempt to improve the world by investing in companies that function in an ethical manner. SRI is frequently described as the attempt to 'do good while doing well.' "
However, modern portfolio theory suggests that SRI will have a small effect on most firms' cost of capital (Beltratti, 2003). An exception is possible if SRI raters identify small niche firms that are otherwise "below the radar screen" of most investors (as in Merton 1987). On the other hand, SRI ratings can have much larger effects if they impact perceptions of consumers, employees, managers, regulators or other stakeholders.

Investors with deontological motives consider it unethical to receive profits from sectors they consider evil or harmful (Rosen, Sandler, and Shani, 1991). Such investors might make these portfolio choices even knowing that their decisions do not raise the cost of capital for firms the investors dislike. For example, the Methodist Church's stock market investments have carefully avoided firms involved in alcohol and gambling.[25]

Finally, the expressive motive for social investment involves investing in firms marked as "responsible" because such choices express the investor's own social responsibility to both the investor and others. Expressive motives for SRI arise because, in our culture, many people perceive that "good" people act "socially responsibly." For example, 80% of Americans call themselves environmentalists. In addition, 75-80% of Americans say they would pay more for environmentally responsible products (although in reality, such products have a far lower share of the market). Thus, it is internally consistent for a good person to invest in a fund that calls itself "socially responsible."

---

[25] Ethical Investment Research Services, "A Brief History of SRI/ Ethical Investment" (http://www.eiris.org/pages/top menu/key facts and figures/history of ethical investment.htm, accessed May 3, 2007).

Appendix 3: Estimation of KLD and Calvert scores

Predicting membership in KLD's Large Cap Social Index (LCS) with KLD subscores
Logit: Results are expressed as dP/dX.

Variables	Coefficients
Generous Giving	0.06
Innovative Giving	-0.08
Support for Housing	0.06
Support for Education	0.13**
Non-U.S. Community Involvement	-0.04
Volunteer Programs Strength	0.1
Investment Controversies	-0.43**
Negative Economic Impact	-0.47**
Tax Disputes Concern	-0.04
Other Community Concern	0.012
Limited Compensation-Strength	0.053
Transparency Strength	-0.02
Political Accountability Strength	0.09
Other Corporate Governance Strength	-0.1
High Compensation-Concern	0.07
Accounting Concern	-0.29
Political Accountability Concern	0.05
Other Corporate Governance Concern	-0.11
CEO-Diversity	0.02
Promotion-Diversity	0.06
Board of Directors-Diversity	0.08*
Family Benefits-Diversity	-0.09
Women/Minority Contracting	0.08
Employment of the Disabled	0.16**
Progressive Gay and Lesbian policies	0.02
Controversies-Diversity Concern	0.01
Non-Representation-Diversity Concern	-0.05
Other Diversity Concern	-0.4**
Union Relations-Strength	0.1
Cash Profit Sharing Strength	0.04
Employment Involvement Strength	0.09*
Strong Retirement Benefits Strength	0.13**
Health and Safety Strength	-0.09
Other Employment Strength	0.13**
Union Relations-Concerns	-0.32*
Safety-Concerns	-0.11
Workforce Reductions-Concern	-0.13
Pension/benefit-Concern	-0.03
Other Employment Concern	-0.02
Beneficial Products and Services-Strength	0.13**
Pollution Prevention-Strength	0.06
Recycling-Strength	0.16**
Alternative Fuels-Strength	0.13**
Other Environment Strength	-0.01
Hazardous Waster-Concern	-0.19
Regulatory Problems-Concern	-0.07
Substantial Emissions-Concern	-0.35**
Climate Change-Concern	-0.04
Other Environment Concern	-0.25
Product Quality-Strength	0.05
R&D/Innovation Strength	0.13**
Benefits to Economically Disadvantaged-Streng	-0.04
Product Safety-Concern	-0.25**
Marketing/Contracting Concern	-0.11*
Antitrust concern	-0.12
Other Product Concern	-0.34**
Burma concern	-0.68**
International Labor Concern	-0.13
Indigenous People concern	-0.35
Other human rights concern	-0.35

N=856 observations
* Significant at 5% level
** Significant at 1% level

For more detail on subscores, see http://www.kld.com/index.html

Predicting Membership in Calvert's Social Fund with Calvert subscores
Logit: Results are expressed as dP/dX

Variables	Coefficients
Environment: Management & Policies, Performance & Impact, Product Lifecycle, Resource Use & Habitat	0.27**
Workplace: Diversity, Labor Relations, Employee Health & Safety	.40**
Business Practices: Corporate Governance, Business Ethics, Product Safety & Impact, and Animal Welfare	.52**
Human Rights: Management & Policies, Performance, Indigenous Peoples' Rights	0.14
Community Relations: Economic Impact, Community Unrest, Philanthropy, Employee Volunteerism, Fair Lending	0.1

N=100 observations
* Significant at 5% level
** Significant at 1% level

For more detail on subscores, see http://www.calvert.com/sri_7889.html and http://www.calvert.com/sri_calvertratings.html
Companies are rated on a scale of 1 (substantially below Calvert standards) to 5 (superior).

Appendix 4: How much do the memberships of top Social Ratings funds intersect? *

Index and [maximum N in index]	Calvert	FTSE4Good Top 100	DJSI	Innovest Elite 17
KLD's LCS (670)	89% of the 493 Calvert members, but only 47% of the 490 Calvert nonmembers, are in the LCS.	79% of the 78 FTSE4Good top 100 are in LCS, and 65% the 513 FTSE nonmembers	DJSI's 60 members and 20 waitlist are both about 65% in LCS, as opposed 60% of the DJSI "others"	47% of the 17 Innovest elite and 67% of the nonelite are in LCS
Calvert Social Index [607]	x	61% of the 77 FTSE4Good top 100 and 51% of the 501 FTSE "others" are members of Calvert	60% of the 57 DJSI members, 55% of the 20 waitlist, and 53% of the 834 others are in Calvert	59% of the 17 Innovest elite and 51% of the 485 non-elite are in Calvert
DJSI top 10% plus 1 in each industry [88]	x	x	42% of the 52 DJSI members are in FTSE4Good top 100, as opposed to 33% of the 15 waitlist firms and 10% of the 503 DJSI "others".	5 of Innovest's 14 elite are in the 75 FTSE4Good, while 11 are in the 317 nonmembers.
FTSE4Good Top 100 [101]	x	x	x	15% of DJSI's 59 members are in Innovest's elite 18. This share is small, but it is higher than the 6% of the 20 DJSI "waitlist" firms and higher yet again than the 2% of DJSI "others".
Innovest global top 100 [17 in U.S.]	x	x	x	x

Appendix 5: Correlations among continuous scores

We examined correlations among 3 continuous scores: KLD Scores (described in the main text), Calvert Scores, and DJSI rankings (for the top 10% plus one firm per industry). We computed Calvert Scores by regressing the 5 Calvert subscores on Calvert membership using a logit equation. As with the KLD score, the Calvert Score is the predicted probability of membership from this regression. The KLD Score estimated based on sub-scores is correlated 0.49 with the Calvert Score. This correlation is close to the estimated tetrachoric correlation based on overlapping membership. Recall the tetrachoric correlation is based on all 1000 members of the universe of firms both KLD and Calvert rated, while the correlation of continuous scores is based on only 100 large firms where Calvert publishes subscores.

Results for the DJSI rating are less encouraging. DJSI ratings are available only for the top 10% of each industry plus one runner-up per industry. As such, the correlations can be reduced by the restriction of range. Within that upper strata, the correlation between DJSI rankings and KLD Scores is only 0.03 (n.s.) (Table 3, col. 1) and between Calvert Scores and DJSI Rankings is only 0.007 (n.s.). These findings provide no evidence that DJSI's top rated firms correlate with KLD's or Calvert's. As with our previous results, adjusting the continuous scores for visible differences in ratings did not raise the correlations. Starting with the continuous scores, it is straightforward to norm the KLD score by industry and to remove the KLD screens that are not in Calvert (to create an index comparable to Calvert's method) or without any screens (to be comparable to DJSI). The correlations are not particularly larger when we adjust the KLD Score to use the explicitly different methods of Calvert, DJSI, and FTSE. (Rows 2 and 3).

How correlated are the continuous Social Ratings?

Correlations between different "styles" of KLD Score and other continuous social ratings.
 KLD Score = Predicted probability of KLD LCS membership (with adjustments noted in rows 2 & 3)

	Firms affected by screens	Industry normed	Calvert Score	DJSI rank
1. KLD style	KLD's screens get Score = 0	no	0.4861	0.031
2. Calvert style	Calvert's screens get Score = 0.	yes	0.4922	
3. DJSI style	No screens	yes		0.0863
N for overlap			100	88

Notes

Calvert Score predicted probability of Calvert Index membership based on 5 subscores and 1 screen(alcohol)?

DJSI Rank is available for the top 10% of each industry (coded with rank = 90-100 percentile) and 22 waitlisted firms (rank coded 88%xx)

Screens for the other indices are proxied with KLD's measures of that screen. For example, Calvert has a screen on alcohol, but due to data limitations we use KLD's screen for alcohol

If not screened out, the company's score is the predicted probability of membership in the KLD LCS based on its 64 KLD subscores. See appendix 1 for the results of this regression.

Industry norming of KLD score is carried out by including industry dummies in the logit equations predicting KLD LCS membership,

but dropping the dummies when predicting membership.

Appendix 6: Do KLD sub-scores predict fewer scandals?
A. Regression Analysis

Dependent Variable: Any Scandal in Year t [a] Independent Variables in Year t-3	Probit Regression: Marginal Effects Reported Standard Errors in Brackets	Conditional logit
Community Strength	-0.069	-0.26
	{.048}	[.232]
Community Concern	0.119	0.533
	{0.102}	[.460]
Diversity Strength	0.067**	0.379***
	{.03}	[.155]
Diversity Concern	0.108*	0.575**
	{0.061}	[.273]
Employee Strength	0.038	0.148
	{.043}	[.187]
Employee Concern	0.08	0.451**
	{0.049}	[.212]
Environmental Strength	-0.062	-0.244
	{0.056}	[.278]
Environmental Concern	0.009	0.143
	{0.036}	[.199]
Product Strength	0.065	0.357
	{0.082}	[.358]
Product Concern	0.065	0.461**
	{0.04}	[.204]
	SE adjusted for clustering by pair	Fixed effect for each pair
	N=428, Psedo R Squared=.0365	N=420, Psedo R Squared=0.0967

a: Firms with scandals are all firms with restatements, firms on the top corporate criminals list, firms on the 100 largest environmental spills list, the Wall Street Journal Perfect Payday list, and other firms on a list collected by the authors that can also be found in KLD DS400.

^^ Firms without scandals were chosen by matching on # employees(logged) in year t-3 and 2 digit industry classification in year t-3(used t-2 when t-3 was unavailable)

^^^ ***, **, * Significant at the 1%, 5%, 10% level

B. Comparing Means of Scandal and Comparison Firms

	Mean (Scandal Firms) (N=218)	Mean (Non-Scandal Firms) (N=210)	Gap	T-Stat
Community Strength	0.25	0.22	0.03	-0.5
Community Concern	0.087	0.048	0.039	-1.63
Diversity Strength	0.7	0.46	0.24	-2.53
Diversity Concern	0.25	0.18	0.07	-1.77
Employee Strength	0.43	0.32	0.11	-1.77
Employee Concern	0.34	0.24	0.1	-2.04
Environmental Strength	0.179	0.186	-0.007	0.15
Environmental Concern	0.35	0.29	0.06	-0.713
Product Strength	0.13	0.1	0.03	-1.2
Product Concern	0.37	0.23	0.14	-2.12
Sum of Strengths	1.7	1.3	0.4	-2.17
Sum of Concerns	1.4	1	0.4	-2.96
Sum of Strengths-Concerns	0.29	0.28	0.01	-0.06