Strategic Management Journal
Strat. Mgmt. J., **37**: 66–85 (2016)
Published online EarlyView in Wiley Online Library (wileyonlinelibrary.com) DOI: 10.1002/smj.2463
Received 29 August 2013; *Final revision received 23 January 2015*

USING ITEM RESPONSE THEORY TO IMPROVE MEASUREMENT IN STRATEGIC MANAGEMENT RESEARCH: AN APPLICATION TO CORPORATE SOCIAL RESPONSIBILITY

ROBERT J. CARROLL,[1] DAVID M. PRIMO,[2] and BRIAN K. RICHTER[3]*
[1] *Department of Political Science, Florida State University, Tallahassee, Florida, U.S.A.*
[2] *Department of Political Science and Simon Business School, University of Rochester, Rochester, New York, U.S.A.*
[3] *Business, Government, & Society Department, McCombs School of Business, University of Texas, Austin, Texas, U.S.A.*

Research summary: *This article uses item response theory (IRT) to advance strategic management research, focusing on an application to corporate social responsibility (CSR). IRT explicitly models firms' and individuals' observable actions in order to measure unobserved, latent characteristics. IRT models have helped researchers improve measures in numerous disciplines. To demonstrate their potential in strategic management, we show how the method improves on a key measure of corporate social responsibility and corporate social performance (CSP), the KLD Index, by creating what we term D-SOCIAL-KLD scores, and associated estimates of their accuracy, from the underlying data. We show, for instance, that firms such as Apple may not be as "good" as previously thought, while firms such as Walmart may perform better than typically believed. We also show that the D-SOCIAL-KLD measure outperforms the KLD Index and factor analysis in predicting new CSR-related activity.*

Managerial summary: *Corporate social responsibility (CSR) continues to grow in importance among the press, political activists, managers, analysts, and investors, yet measurement techniques have not kept up. We show that the most common approach for measuring CSR—adding up observable traits—is fundamentally flawed, even if these traits accurately capture CSR-related behavior. We introduce an improved measurement technique that treats these traits as test questions that are differentially weighted, so that "hard" CSR activities affect a company's score more than "easy" CSR activities. This approach produces a measure that offers a more reliable comparison of firms than standard measures. Our approach has a number of additional advantages, including differentiating firms that receive identical scores on an additive scale and accounting for how CSR-related behavior has evolved over time. Anybody who cares about CSR should consider using our measure (available at www.socialscores.org) as the basis for analyzing firms' CSR.*

INTRODUCTION

The core challenge to measurement in strategic management contexts is that, unlike in the physical sciences, the firm-level and individual-level characteristics we would like to measure are often inherently impossible to observe directly (Godfrey and Hill, 1995). For example, how can we determine, in an objective manner, how well-governed (e.g., Aguilera and Jackson, 2003; Daily, Dalton, and Cannella, 2003; Shleifer and Vishny, 1997), entrepreneurial (e.g., Covin and Slevin, 1991; Lumpkin and Dess, 1996), or socially responsible (e.g., Carroll, 1979) a given firm really is? The

Keywords: measurement; item response theory; Bayesian estimation; corporate social responsibility; corporate social performance
*Correspondence to: Brian K. Richter, Mailing: 2110 Speedway, B6500, CBA 5.250, Austin, TX 78712-0177. E-mail: brian.richter@mccombs.utexas.edu

challenge is so great that poor measurement has been called "one of the most serious threats to strategic management research" (Boyd, Gove, and Hitt, 2005). This article shows how researchers can use item response theory (IRT) modeling to improve measurement; we demonstrate the usefulness of IRT with an application to corporate social responsibility/performance (CSR/CSP).

Researchers often construct measures built from multiple observable proxies using either additive indices or data reduction techniques such as factor analysis (Boyd *et al.*, 2005). These approaches have several benefits compared to the use of a single proxy; for instance, they make use of more information and reduce measurement error that might arise from one noisy signal. However, they also have serious drawbacks. The implicit assumption underlying the construction of additive indices, for instance, is that each observable is an equally good proxy of the underlying attribute we hope to measure. This, of course, is a strong assumption that is difficult to justify theoretically, yet additive indices are used in a variety of contexts, including CSR/CSP (the focus of this article) and the "G-index," which is used to measure the quality of corporate governance (Aguilera and Desender, 2012). While an improvement over additive indices, scales based on data reduction techniques like factor analysis—the firm-level "entrepreneurial orientation (EO) scale" (Lyon, Lumpkin, and Dess, 2000) being a prominent example—are not as flexible as the IRT approach we introduce in this article.

IRT MODELS

Item response theory (IRT) models can improve on existing "state of the art" measurement techniques by generating measures of latent characteristics based on a richer, theory-driven understanding of how these characteristics are reflected in proxies. In doing so, IRT models enable the researcher to assess important questions. Are differences between individuals and firms in traditional measures of latent characteristics real or due to systematic measurement error (which can be estimated for IRT-based measures)? How do individual firms and groups of firms change over time? Are some items in an index better/worse at distinguishing among firms, and if so, by how much?

The data inputted into an IRT model for estimation of latent traits may be a set of responses to a series of questions or a set of other observed measures, such as whether various behaviors occurred or did not occur.[1] Extrapolating from an education setting, these observables can be thought of as answers to test questions, following Thurstone (1925), who had the insight that students of varying ability levels respond differently to various test questions, which themselves vary in how well they measure ability (Bock, 1997). Hence, IRT models simultaneously assess both the test questions and the test takers.

We focus here on a basic two-parameter model for binary (e.g., yes-no; absent-present; 0–1; correct-incorrect) data. IRT models can also accommodate ordinal responses (e.g., a rating on a scale of 1–5) and additional parameters. In the article's conclusion, we discuss how management researchers can take advantage of this flexibility.

The basic model takes the following form: $\Pr(y_{i,j} = 1 \mid \rho_i, \alpha_j, \beta_j) = F(-\alpha_j + \beta_j \rho_i)$. The i subscript refers to individual respondents, while the j subscript refers to the items used to assess those respondents. $F(\cdot)$ is typically the logistic or standard normal function, making this formula similar to a logit or probit model when working with binary data (Hoetker, 2007); a key difference between applications of those techniques and IRT models, however, is that in IRT there is typically no independent variable with observed data (i.e., x_i); rather, it is replaced by the ρ_i term representing ability (or another latent trait) that the researcher wishes to estimate. The outputs of a basic two-parameter model are estimates of the latent trait for each individual in the dataset (ρ_i), along with estimates for how difficult each item is (α_j) and how well each item discriminates among individuals (β_j). Using a test analogy, α_j addresses the question "Holding ability fixed, how likely is a student to get question j correct?", and β_j addresses the question "How well does question j help distinguish between students of different ability levels?"; in other words, do individuals with high ability and low ability (i.e., high and low ρ_i s) differ in the probability they will get a question correct?

IRT models have deep roots in psychology (Lord and Novick, 1968; Rasch, 1960; Reise and Waller,

[1] The discussion in this section draws from Johnson and Albert (1999), and Fox (2010).

DOI: 10.1002/smj

2009) and have made inroads into many disciplines, including economics (e.g., Høyland, Moene, and Willumsen, 2012) and medicine/public health (Das and Hammer, 2005; Faye *et al.*, 2011; Hays and Lipscomb, 2007; Hedeker, Mermelstein, and Flay, 2006). The closest analog to the IRT analysis in this article, however, comes from political science, given parallels in the structure of data on observable behavior in political science and management. The classic use of IRT models in political science is estimating legislators' ideology (or "ideal points") on a left-right continuum (Clinton, Jackman, and Rivers, 2004; Jackman, 2000; Londregan, 2000; Poole and Rosenthal, 1991) to improve on crude proxies like party affiliation (e.g., Bonardi, Holburn, and Vanden Bergh, 2006; Vanden Bergh and Holburn, 2007).

MEASURING CORPORATE SOCIAL RESPONSIBILITY ACTIVITY

Corporate social responsibility is undoubtedly an important topic for strategic management researchers today: the term, or one of its close analogs, appeared in nearly 50 percent of *Strategic Management Journal* issues over the five-year period from 2008 to 2012. It is also an area fraught with interrelated conceptual and measurement issues.[2]

The CSR construct is a complicated one that may be manifested in a number of different behaviors, depending on firm-specific factors and competing definitions (Carroll, 1979, 1999, 2009; Dahlsrud, 2008). To some, the term *CSR* itself is problematic because the construct "responsibility" reflects value structures that vary from firm to firm: "The term is a brilliant one; it means something, but not always the same thing, to everybody. To some it conveys the idea of legal responsibility or liability; to others, it means socially responsible *behavior* in an ethical sense," and so on (Votaw, 1972: 25). More recently, Wood (1991: 699) has argued that because CSR content will "vary somewhat from company to company," measurement should focus on social outcomes.

It's not surprising, then, that early data-driven work related to CSR "was plagued with measurement problems, because few good measures existed for the multidimensional construct" and researchers tended "to select a single item as a proxy" (Surroca, Tribó, and Waddock, 2010). In response to these challenges, Frederick (1994) argued for sidestepping the CSR construct altogether by limiting interpretations of findings to "narrower and more technical" definitions labeled corporate social performance (CSP). Since then, numerous competing perspectives on the distinction between CSR and CSP have emerged (e.g., compare Barnett, 2007; Baron, 2001) . For ease of exposition, in what follows, we use the terms *CSR* or *CSR-related activity*, but we just as easily could have used the term *CSP*.

Despite all of these challenges, things began to look up for the measurement of corporate responses to the CSR construct when Waddock and Graves (1997) introduced the KLD STATS (Statistical Tools for Analyzing Trends in Social and Environmental Performance) dataset to academic researchers. The KLD STATS data were the first to capture a large set of firm-specific actions related to the CSR construct across a large number of categories and for a broad cross-section of firms over several years (MSCI ESG Research, 2012).

The KLD Index can be constructed for a given firm in a given year by summing up a large number of binary "strength" indicators and subtracting out a large number of binary "concern" indicators that KLD researchers code. The KLD STATS dataset includes more than 80 binary indicators of whether or not a given firm meets or does not meet an objective, "observed/not observed" behavioral criterion across eight broad categories related to CSR including the environment, community, human rights, employee relations, diversity, product attributes, governance, and involvement in controversial business issues. KLD refers to some indicators as "strengths," which proxy social responsibility, and other indicators as "concerns," which proxy social irresponsibility.

The KLD dataset is "the *de facto* research standard" (Waddock, 2003) in this literature, but an entire literature also has emerged where the primary purpose is to critique or assess the validity of the KLD Index, often on the same grounds as those for other equally-weighted indices alluded to in the introduction. Articles of this sort include Sharfman (1996), Griffin and Mahon (1997), Rowley and Berman (2000), Entine (2003), Graafland,

[2] For background on the CSR literature, it is worth looking at one of the numerous literature reviews (e.g., Aguinis and Glavas, 2012; deBakker, Groenewegen, and Den Hond, 2005; Griffin and Mahon, 1997; Kitzmueller and Shimshack, 2012; Margolis and Walsh, 2003; Orlitzky, Siegel, and Waldman, 2011) or meta-analyses (e.g., Margolis, Elfenbein, and Walsh, 2009; Orlitzky, Schmidt, and Rynes, 2003).

DOI: 10.1002/smj

Eijffinger, and Smid (2004), Mattingly and Berman (2006), Sharfman and Fernando (2008), Chatterji, Levine, and Toffel (2009), Delmas and Blass (2010), Walls, Phan, and Berrone (2011), and Delmas, Etzion, and Nairn-Birch (2013). We turn back to these critiques after presenting our new measure: the D-SOCIAL-KLD score, which stands for **D**ynamic **S**tudy **O**f **C**orporate Social Responsibility/Performance with **I**RT **A**na**L**ytics, as applied to the KLD data.

APPLICATION: OUR MODEL AND DATA

In this section, we introduce the key theoretical elements of our IRT model for CSR-related activity and discuss its translation to the estimation itself.

Theoretical model

We adopt a simple, but powerful, theoretical conception of corporate decision making in constructing our IRT model. More precisely, drawing from the theoretical framework in Clinton *et al.* (2004), we devise a model focusing on the utility, or benefit, that a firm receives from adopting (or not adopting) a particular CSR-related policy (e.g., a recycling program). Let $u^d_{i,j,t}$ represent the utility that firm i obtains from making decision d on observable CSR policy j in time period t. Firm i's utility is a function of its underlying, latent level of CSR ($\rho_{i,t}$), the level of CSR/CSP reflected in pursuing CSR policy j for all firms ($\tau^d_{j,t}$), and an error component ($\xi^d_{i,j,t}$). The utility is modeled as a simple quadratic loss function: $u^d_{i,j,t} = -\left|\rho_{i,t} - \tau^d_{j,t}\right|^2 + \xi^d_{i,j,t}$. Such loss functions are standard in the literature, as they are easy to work with and tap into the natural sense of "distance" that underlie spatial models. That is, the utility for adopting a pro-CSR policy is a function of how "far" the resulting CSR policy is from the firm's unobservable level of CSR, plus an error term (which will be important for estimation) reflecting idiosyncratic factors that may also play a role in the firm's decision. Similarly, the utility from not adopting the policy is a function of whether the nonadoption is consistent with the firm's underlying responsibility. It is straightforward to adapt the logic for CSR "concerns" instead of "strengths."

The firm chooses to adopt a policy (A) rather than to reject it (R) if it receives a higher utility from adoption than rejection (i.e., if its net benefit of adoption is positive). Let $z_{i,j,t}$ represent firm i's net benefit for choosing to adopt a policy on observable j in time period t. This can be represented as $z_{i,j,t} = u^A_{i,j,t} - u^R_{i,j,t}$. We can substitute the formulas above into this equation and simplify:

$$\begin{aligned} z_{i,j,t} &= u^A_{i,j,t} - u^R_{i,j,t} \\ &= -\left|\rho_{i,t} - \tau^A_{j,t}\right|^2 + \xi^A_{i,j,t} + \left|\rho_{i,t} - \tau^R_{j,t}\right|^2 - \xi^R_{i,j,t} \\ &= \left(\tau^R_{j,t}\tau^R_{j,t} - \tau^A_{j,t}\tau^A_{j,t}\right) + 2\left(\tau^A_{j,t} - \tau^R_{j,t}\right)\rho_{i,t} \\ &\quad + \left(\xi^A_{i,j,t} - \xi^R_{i,j,t}\right) \\ &= \alpha_{j,t} + \beta_{j,t}\rho_{i,t} + \varepsilon_{i,j,t}. \end{aligned}$$

The simplification from τ terms to α and β terms is necessary for estimation, but it also is true that α, β, and ρ represent substantively meaningful quantities. This formula, in fact, shares the same structure as the two-item IRT model equation presented earlier, though now it is necessary to discuss these parameters in the context of our current application. Using the language of the IRT literature, $\alpha_{j,t}$ is the *difficulty* parameter for adopting policy j in time period t. This terminology should not be taken literally. Instead, $\alpha_{j,t}$ can be thought of as the likelihood that a firm adopts policy j, given a particular level of CSR. In other words, as $\alpha_{j,t}$ increases, all firms are more likely to adopt policy j at time t, although the magnitude of the effect will typically depend on the firm's CSR level due to nonlinearities in the probability model used to generate the estimates. $\beta_{j,t}$ is the *discrimination* parameter for adopting policy j in time period t. If $\beta_{j,t}$ is positive, then more socially responsible firms are more likely to adopt policy j; if it is negative, then more socially responsible firms are less likely to adopt j. Thus, $\alpha_{j,t}$ and $\beta_{j,t}$ tell us about *policy-specific* characteristics. Finally, $\rho_{i,t}$, which represents the underlying *responsibility* for firm i in time period t, is the model's sole assessment of the firm's latent qualities given the policy-specific qualities. $\rho_{i,t}$ is our primary quantity of interest in this article.

The goal is to estimate all three sets of parameters using the actual policy decisions themselves. Put together, this approach allows the data to help the analyst assess *how* particular strengths and concerns map into CSR. Just as a "liberal" legislator is one that follows a particular pattern of "yea"

DOI: 10.1002/smj

and "nay" votes depending on the matter at hand, so too is a "responsible" firm one that follows a particular pattern of corporate decisions or policies. But our approach also allows the analyst to learn about the nature of the policies themselves: if a set of "responsible" firms all adopt a particular policy, then we would think that the policy is a strength rather than a concern (and likewise for irresponsible firms and concerns). The end result, then, is a *dimension* that places firms and policies along a single responsibility line. The dimension separates the responsible from the irresponsible, the strength from the concern.

The Bayesian approach

To this point, nothing about our model necessitates a particular kind of estimation strategy; we have only specified a theoretical model of how firms make decisions on which CSR-related policies to adopt, given some unobservable level of CSR.[3] Building on the work of Martin and Quinn (2002), we adopt a Bayesian mode of inference for both theoretical and pragmatic reasons.

The Bayesian approach, unlike frequentist approaches such as maximum likelihood estimation (e.g., probit), treats the unknown parameters as random variables (i.e., variables that can take on different values, each of which is assigned an associated probability). The Bayesian approach starts with the researcher's best guess (or "prior") about the distribution of these parameters and uses simulations based on observed data to update this guess and produce a "posterior distribution" of the parameters of interest, which in turn, can be used to obtain meaningful results such as a point estimate and confidence bands. Bayesian methods often require computationally intensive tools, most notably Markov chain Monte Carlo (MCMC) algorithms, and our approach is detailed in Appendix S1.

All modeling requires assumptions, and Bayesian models are very flexible in this regard. For instance, because our dataset has a time component, we must make some assumptions about dynamics with both theory and tractability in mind. For the responsibility measures, we assume that the scores are drawn from a normal distribution with mean equal to the previous year's score and variance equal to $\Delta_{\rho_{i,t}}$, which is estimated as part of the model and dictates how closely information from the previous period relates to information in the current period. For the difficulty and discrimination terms, we do *not* model dynamic effects in policy-specific attributes. Instead, to make the model more tractable, we treat each observable as a "new case" in each year.

Bayesian approaches have several other practical benefits beyond the incorporation of dynamics. First, they can more readily be used with large datasets. Second, the estimation of simulated distributions means that we can get a nuanced picture of the accuracy of our estimates. Third, analysts can make use of "priors" to incorporate additional theoretical information into the model. In sum, the benefits of Bayesianism rest on the explicit simulation of entire posterior distributions—thus giving more relevant values of uncertainty for future analysts—and in the ability to estimate all of them together feasibly, even handling missing data with ease.

Data

We utilize the KLD STATS data described above from their inception in 1991 through 2012. The KLD data include a wide variety of indicators, more than 80 per year, each measured dichotomously and coded 1 if the indicator is adopted and 0 otherwise. Across 22 years, we observe a total of 1,610 indicators. On the firm side, the KLD data have included more and more firms over time. From 1991 to 2000, they covered only those firms in the S&P 500 and the Domini 400 Social Index (approximately 650 firms per year, in total). Of course, firms entered and exited those indices over time, so it was not the *same* 650 firms per year. In 2001, KLD expanded its coverage to include all firms that were among the 1,000 largest in the United States, taking the total up to roughly 1,100 per year. In 2002, KLD expanded its coverage further, adding firms in the Large Cap Social Index, with no net change in the total number of firms. From 2003 onward, the data have also included firms from the 2000 Small Cap Index and the Broad Market Social Index, bringing the total to around 3,100 firms per year. All told, our data include 5,784 unique firms over 22 years.

The final data matrix, then, includes $1,610 \times 5,784 = 9,312,240$ unique data cells. Of course, not all of these cells include actual data.

[3] This section relies on background information in Gelman *et al.* (2003) and Fox 2010.

DOI: 10.1002/smj

Not all firms are in the data for all years. Moreover, not all indicators in the KLD data are available for all firms in all years. For the purposes of including as much relevant information as possible, we estimate the model on the entire KLD dataset. Our data matrix, then, includes many missing observations: all told, approximately 70 percent of the observations in the data matrix are missing, leaving us with 2,749,140 actual observations.

Clearly, the missing data issue looms large and has to be handled carefully. We treat missing data in the following way. If a firm is not included in the data up to a particular year, then that firm is not included in the estimation for that year, and thus, has no effect on the estimates. For example, firms that were not in the S&P or Domini indices through the 1990s are not included in the data for those years, and so their D-SOCIAL-KLD scores are not estimated until they do enter the data. Once a firm enters the data, it is treated as part of the population—regardless of whether it is observed (in general or for a particular observed indicator) in a given year—so long as it is again included in the data at some point. For example, Exxon and Mobil are estimated as independent firms through 1999, and then are not estimated thereafter; instead, the single firm ExxonMobil enters the data in 2000 and is estimated through the rest of the time frame.

Our application is novel not only substantively in its focus on CSR, but also methodologically. This is a massively large dataset, and even a few years ago, limits on computational power would have made this estimation infeasible. Indeed, even the *results* are massive: We simulate values for each of the 1,610 α terms and 1,610 β terms (a pair for each observed policy-year estimated) along with the values for each of the 40,505 ρ terms (one for each firm-year estimated). Our final result is a simulation of the complete joint posterior distribution of all 43,725 parameters in the model. We provide 2,500 draws from the joint posterior distribution, meaning that the final data matrix has 109,312,500 unique elements. Below, we present only a small slice of the results from our estimation, focusing on ρ, the unobservable level of CSR, in the name of demonstrating IRT's utility not only in an application to improving the measurement of CSR, but also to improving the measurement of other unobservable constructs in strategic management contexts. We leave potentially interesting discussions about the items themselves (through an analysis of α and β) to future work.

To help researchers adapt IRT as a tool to improve measurement in this and other strategic management contexts, we will be making replication code available at http://www.socialscores.org, where we will also share the firm-year scores we produce in this article.

APPLICATION: RESULTS

The IRT model takes a very large data matrix full of binary responses and missing observations and produces D-SOCIAL-KLD scores linking observations from multiple years. While the overall distribution of D-SOCIAL-KLD scores is roughly centered around zero, the zero point itself has no innate meaning; in the language of statistics, these are interval data, not ratio data. What matters in these scores, just as with KLD Index values, is how firms do relative to one another, and that is our focus in what follows.

Explicitly accounting for measurement error in CSR

We begin our presentation of results by graphing the D-SOCIAL-KLD scores we estimated for all firms in 1991 in (a) and in 2005 in (b) of Figure 1.

We choose to display 1991 because it is the year KLD began rating firms and because it is the year with the fewest firms (647)—making an explanation more straightforward than for other years. While difficult to see, even in 1991, given the size of our dataset, there is a dot (and line) representing each of the 647 firms that KLD covers in its first year. For example, the bottom-most observation in 1991 corresponds with Golden West Financial, while the highest score goes to DuPont, which is consistent with what Delmas and Blass (2010) find in a detailed case analysis and thought experiment applied to 15 firms in the chemicals industry.

The lines for each firm, which are perhaps the most notable feature of this figure, help us demonstrate the power of Bayesian approaches to IRT estimation—as they represent 05–95 inter-percentile ranges (which are analogous to a confidence interval in frequentist statistics). In 1991, there is substantial overlap in the inter-percentile ranges for many firms, especially in the middle of the pack—in fact, fewer than 30 percent of firms can be said to have a latent level of CSR greater than the median and fewer than

DOI: 10.1002/smj

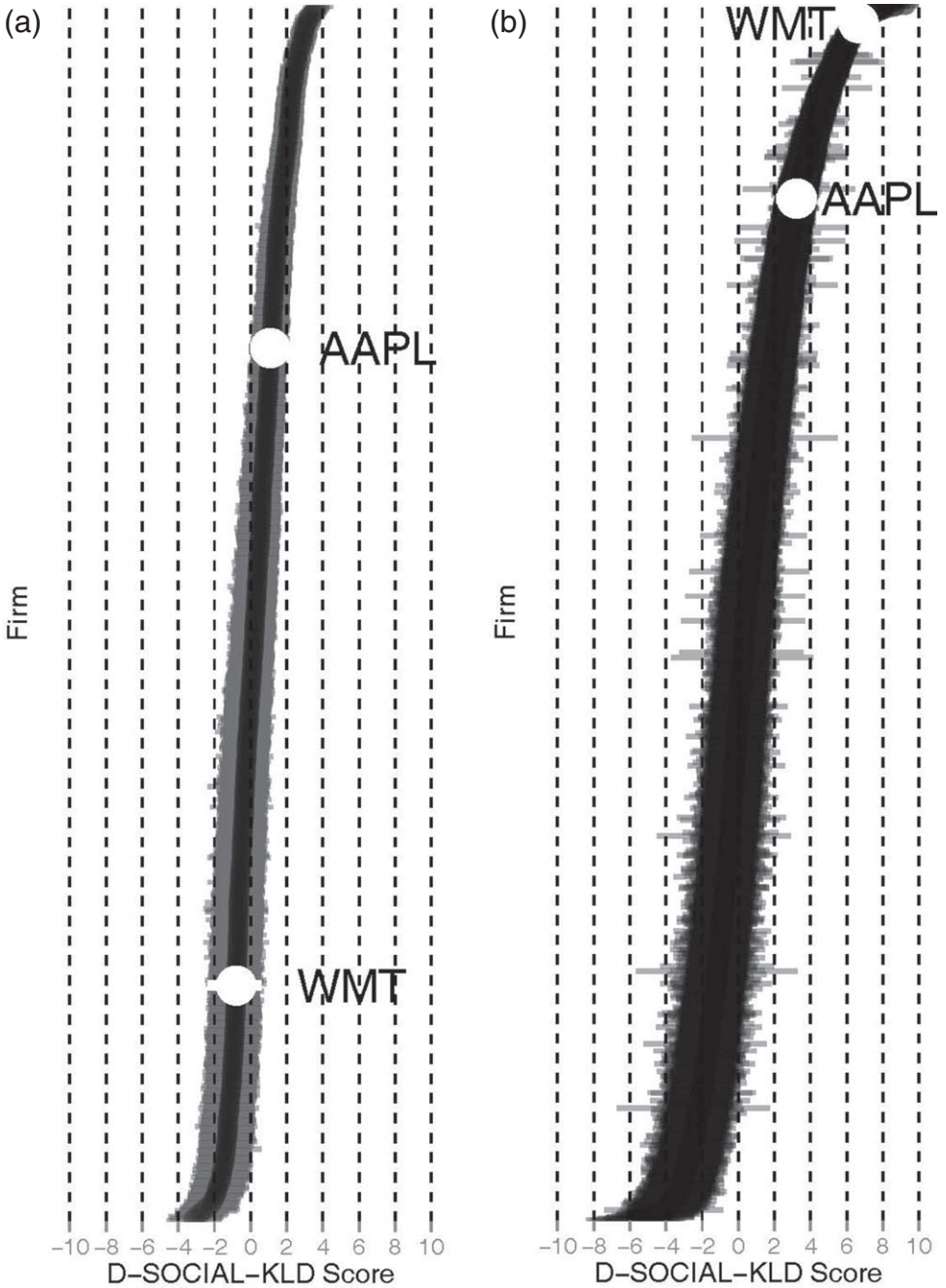

Figure 1. All firms in (a) 1991 and (b) 2005

five percent can be said to have a latent level of CSR less than the median. This overlap indicates that it is difficult to distinguish between the levels of CSR for 65 percent of firms in 1991.

Moreover, the firms toward the top of the graph tend to be simulated with more precision than the firms toward the bottom. This occurs because many of the firms toward the top have been covered by KLD in more years than those that fall toward the bottom, many of which exit the S&P 500 and Domini indices in the 1990s.

This takes us to the first two lessons we glean from the Bayesian estimation of our IRT model. First, *firm-to-firm comparisons of CSR/CSP using*

DOI: 10.1002/smj

measures based on the KLD data should proceed with caution unless the differences in any measure are sufficiently large. Second, *researchers should explicitly account for measurement error* when incorporating KLD-based measures into their empirical analyses. While these points flow directly from the Bayesian application, they have clear implications for the use of other additive indices in strategic management research where measurement error is not explicitly quantified and where there are even fewer observable indicators of latent traits than the 80 here.

Also note that, in general, the results look something like the cumulative density function (CDF) for a normally distributed variable. This basic pattern holds across all years, although the spread increases over time. For instance, in 2005, shown in (b), rather than the dots sitting nearly vertically on top of each other as in (a), they begin to separate from each other with more firms further to the right or to the left of zero. Overall, this change in the underlying distribution of firms' latent CSR levels over time allows us to make more nuanced comparative statements about firms in later years, despite the cautionary point we made above.

To illustrate this, we have labeled Walmart (WMT) and Apple (AAPL) in Figure 1. Looking at the size of their respective inter-percentile ranges in 1991, we cannot confidently say that Walmart's latent level of CSR, despite falling so much lower in the relative distribution, is distinguishable from Apple's in that year. Nevertheless, by 2005, despite the firms falling closer together in a distribution that incorporates a larger number of firms, we can say, with confidence, that Walmart has a higher latent level of CSR than Apple, contrary to what an additive KLD Index (and the conventional wisdom) indicates, and perhaps more consistent with the actual behavior at these firms. For instance, Walmart in 2005 was demonstrating exceptional levels of social responsibility in leveraging its supply chain to aid Hurricane Katrina victims (Diermeier, 2011; Muller and Kräussl, 2011). Meanwhile, Apple was beginning to engage in activities many would argue are socially irresponsible (Amaeshi, Osuji, and Nnodim, 2008; Christensen and Murphy, 2004; Dowling, 2014)—namely, contracting with a Chinese supplier, Foxconn, that had questionable labor practices (Duhigg and Barboza, 2012), and developing a strategy to aggressively avoid paying taxes in the United States (Duhigg and Kocieniewski, 2012). This brings us to the next point the results of our estimation help us illustrate: the ability to measure changes over time.

Observing changes in the levels of CSR over time

One of the greatest strengths of our approach is that we model firm behavior over time in a single space that accounts for dynamic behavior. This allows us to make comparisons within firms, or groups of firms over time, which, technically, we would not be able to do if we had re-estimated a static IRT model in each annual cross-section.

To highlight the explicit incorporation of time in our model, we present the D-SOCIAL-KLD scores of selected major firms over time in Figure 2.

The solid black lines in Figure 2 illustrate our D-SOCIAL-KLD scores, while the gray shaded areas represent confidence bands for them. Dashed black lines in Figure 2 illustrate KLD Index values. We caution readers that the values of the two measures are not directly comparable given different underlying scales (despite both sharing a median near zero). Nevertheless, the trends in our D-SOCIAL-KLD scores and in KLD Index values can be compared—and we observe some meaningful differences on that front when we do so. The figure demonstrates that many notable firms exhibit marked improvements over time in our D-SOCIAL-KLD scores, while the same is not necessarily true for KLD Index values—bringing into question the validity of the KLD Index when considering the actual circumstances at many of these firms. With respect to our D-SOCIAL-KLD scores, there is also quite a bit of heterogeneity in time trends among firms.

We start our analysis with Walmart, which we discussed in reference to Figure 1 above. Walmart has dramatically increased its level of CSR over time as measured by our D-SOCIAL-KLD score: The firm begins from a very low D-SOCIAL-KLD score—less than 0, which is below the overall median—in 1991, and ends up with one of the highest scores by 2012. Notable, also, is that one of Walmart's primary competitors, Target, begins with a much higher D-SOCIAL-KLD score in 1991, and like Walmart, shows improvement over time; however, the pace of improvement is far less dramatic, and so by 2012, Walmart's performance on our D-SOCIAL-KLD score exceeds Target's (with a .87 probability in the full simulated distribution).

DOI: 10.1002/smj

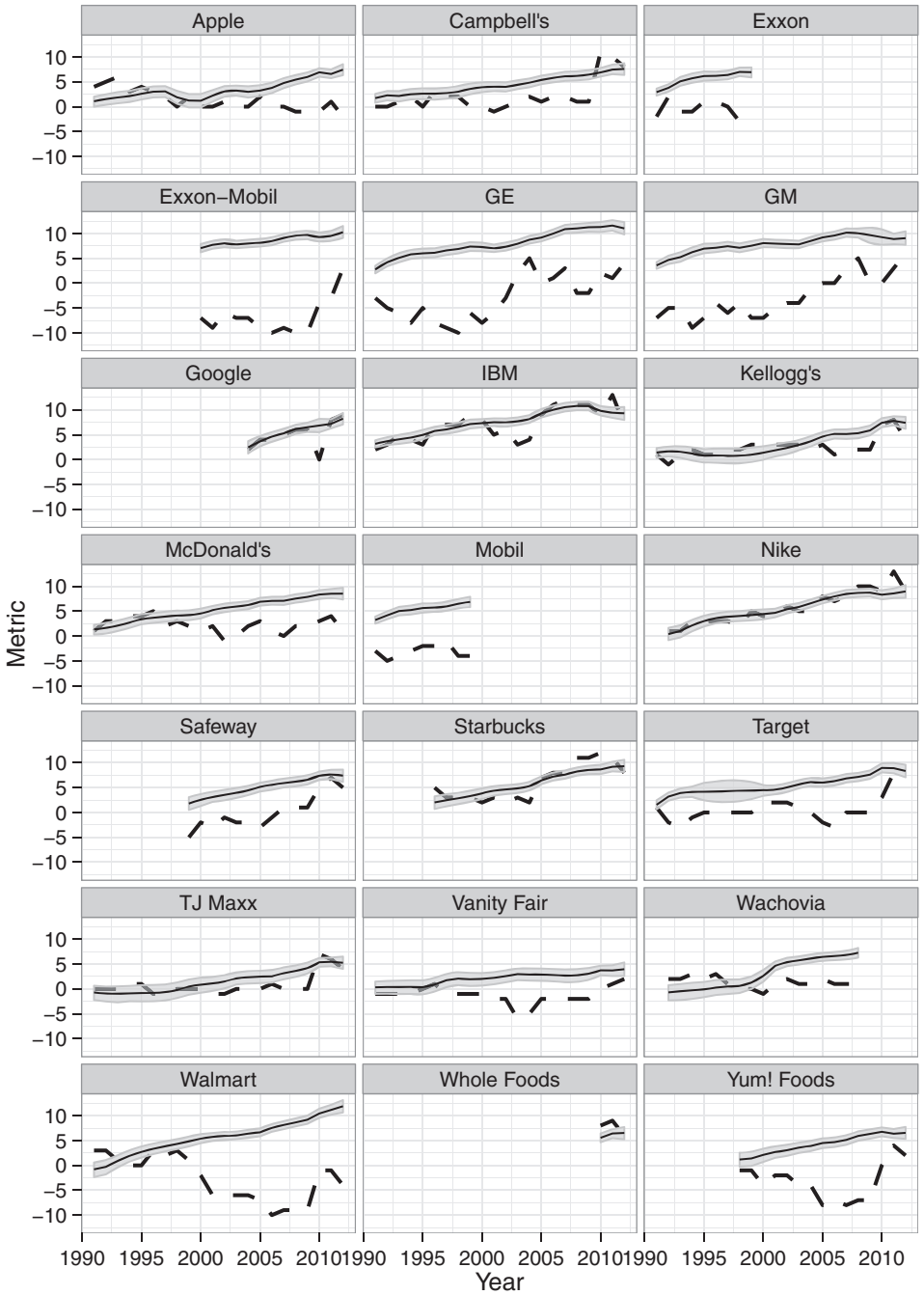

Figure 2. Select firms over time. Bayesian D-SOCIAL-KLD score shown with solid line with confidence interval. KLD Index shown with dashed line

Importantly, had we looked at the KLD Index values alone, we would have come to a very different conclusion when comparing the two companies, as that measure shows an upward trend for Target and a downward trend for Walmart—the latter of which is particularly hard to reconcile with reality given Walmart's recent efforts to be a better corporate citizen (Diermeier, 2011), and calling into question the KLD Index values for these firms.

The positive trend for Walmart and Target in the D-SOCIAL-KLD scores is not necessarily the case for all retailers. We include in the figure

DOI: 10.1002/smj

two notable clothing retailers for bargain-minded shoppers: TJ Maxx and Vanity Fair. Both have very low D-SOCIAL-KLD scores early on and both show relatively small improvement over time in this measure, such that their earlier selves are hardly distinguishable from their later selves when accounting for the widths of the confidence bands. The trends in both retailers' KLD Index values are also relatively flat, although it is harder to say anything about the level of uncertainty in these trends given the lack of error bands.

While many prominent firms show improvement over time, the time trends are not always monotonic. Consider Kellogg's and Apple, both of which demonstrate a general improvement over time despite the occasional downturn. In the case of Kellogg's, the downturn is slow and gradual, whereas Apple's shifts over time are much more sudden. The downward shifts in the D-SOCIAL-KLD score at Apple correspond to periods when founder and sometimes CEO Steve Jobs returns to the firm from temporary hiatuses—consistent with theories about top management driving CSR (up or down) (e.g., Hemingway and Maclagan, 2004; Hong and Minor, 2013).

Another trend to note from this view of the D-SOCIAL-KLD scores is that many firms, particularly those at the high end of the spectrum and also industry leaders like IBM and GM, demonstrate slight downturns toward the end of the data's time span (i.e., in the 2009–2012 period). This suggests, consistent with theory, that CSR may follow economic cycles and be more readily implemented in earnest when firms have slack resources (Campbell, 2007; Hong, Kubik, and Scheinkman, 2012).

We find that large oil companies behave quite similarly to other large firms. As an interesting case, we present results for Exxon and Mobil, which in turn, merge to become ExxonMobil. As independent firms, Exxon and Mobil (and many other similar firms) had nearly identical scores over time, and their merged descendant took up *precisely* where they left off.

We also consider some newer firms with strong reputations as they enter the data. For example, Starbucks enters the data in the late 1990s, and Google does the same in the mid-2000s. Both of these firms begin with average-to-low scores that then improve quickly over time. In contrast, a very new entrant like Whole Foods begins from a much higher starting point. This suggests that new firms may have a more complicated environment to consider in their early growth phases.

Given the overall upward trend for the firms we consider in Figure 2, one might reasonably ask whether this is the case in general. To that end, we plot the median D-SOCIAL-KLD score over time in Figure 3(a). We also plot the median of the KLD Index over time in Figure 3(b).

The overall median in each year is depicted with the solid black line. Prior to KLD expanding its coverage in 2001 to include firms outside the S&P 500 and outside the Domini Social Index, we see in 3(a) that the median firm's D-SOCIAL-KLD score was on the rise. Were we to consider all firms' D-SOCIAL-KLD scores after 2001 in 3(a), we would infer that firms generally became less socially responsible. On further examination, however, S&P 500 and Domini Social Index firms after 2001—depicted with the black dashed line—continued the upward trend, whereas the relatively smaller firms that entered the data in 2001—depicted with the gray dashed line—demonstrated much lower levels of CSR, bringing the overall median downward. Interestingly, these smaller firms, on the whole, persisted at around the same median score through the remainder of the time period.

When we look at the KLD Index data in 3(b) over the same time period, we do not find the same trends. In the KLD Index, all firms, including those in the S&P 500 and Domini Social Index, trend flat over time—which would be inconsistent with the literature on how firms respond to social movements such as CSR and activist demands (Baron, 2001; Baron and Diermeier, 2007; Eesley and Lenox, 2006; Reid and Toffel, 2009). This finding in our D-SOCIAL-KLD score—but not in the KLD Index—might also speak to the claim that large firms are generally less financially constrained than small firms, and hence, more free to spend on CSR initiatives (Hong *et al.*, 2012). The general upward trend for the S&P 500/Domini firms featured in Figure 2—even during recessions—also jives with the view that CSR has over time become viewed by managers as a necessity.

Developing a more nuanced understanding of underlying CSR policies

The differences between the KLD Index and the D-SOCIAL-KLD score require further probing, given the several ways we have already seen them

DOI: 10.1002/smj

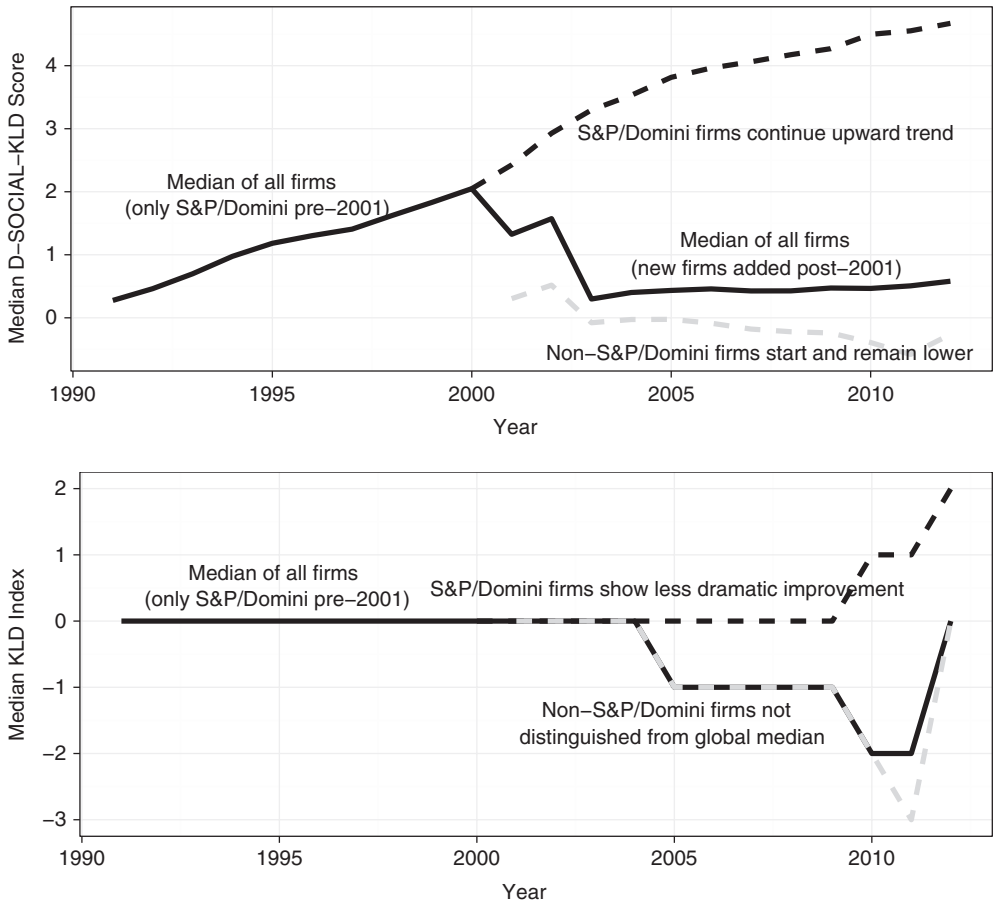

Figure 3. Median scores over time: (a) D-SOCIAL-KLD score medians and (b) KLD Index medians

differ. In Figure 4, we create a scatterplot with the KLD Index on the horizontal axis and the D-SOCIAL-KLD score on the vertical axis.

Each dot in the figure represents a firm-year observation comparing how the D-SOCIAL-KLD scores measure up against the KLD Index values. For emphasis, earlier time points are depicted with darker dots. We highlight overall trends in solid black and include a diagonal dashed line that approximates a one-to-one correspondence between the two measures. In fact, the overall correlation between the KLD Index and our measure is only 0.195, and only in the range from zero and up do the KLD Index values roughly track the D-SOCIAL-KLD scores.

Partially because our D-SOCIAL-KLD score is a continuous measure rather than an ordinal one, there is an enormous amount of heterogeneity among firms with the same KLD Index value. Consider those observations with a KLD Index value of 0, of which there are 10,894. On the surface, it seems odd that over 25 percent of firm-year measures could be identical. It is even odder to think of these firms as being equivalent since there are multiple ways to get to 0; different numbers of different strengths could be summed up and different numbers of different concerns could be subtracted out to reach the same KLD Index value of 0 (e.g., Kotchen and Moon, 2012; Minor and Morgan, 2011; Strike, Gao, and Bansal, 2006).

After all, how similar could latent CSR be at Saul Centers—a real estate management firm that operates around 30 neighborhood shopping centers—and at Ford—one of the world's largest companies—despite both having a KLD Index value of 0? Our D-SOCIAL-KLD scores suggest that, indeed, there are substantial differences between these two firms as they each have very different underlying levels of CSR. Saul Centers has the lowest D-SOCIAL-KLD score (approximately

DOI: 10.1002/smj

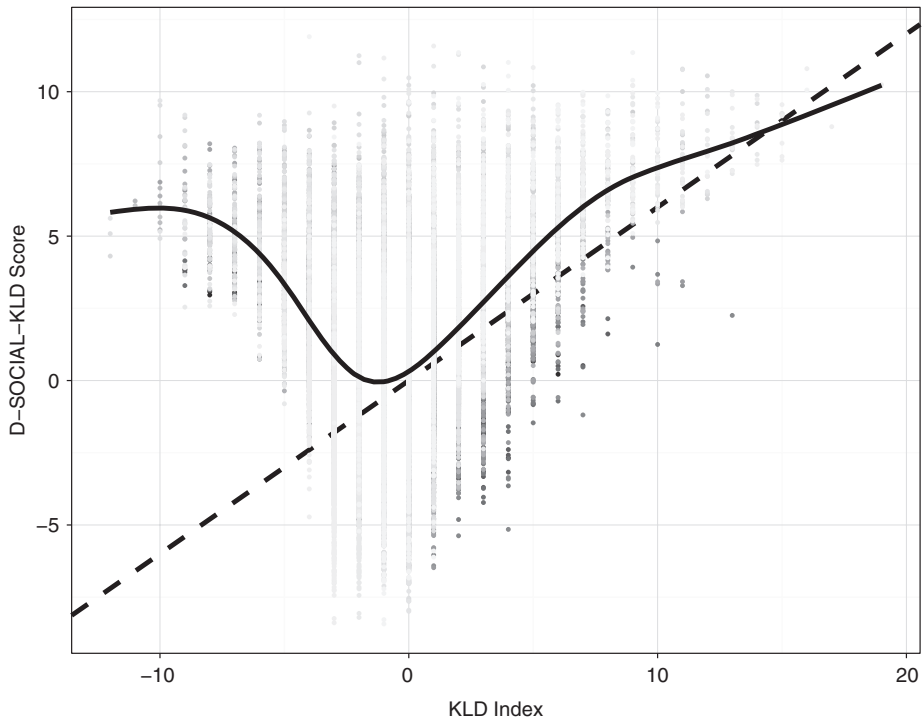

Figure 4. KLD Index versus D-SOCIAL-KLD score by time (earlier timepoints are darker)

-8) among the KLD Index zeroes, whereas Ford has the highest D-SOCIAL-KLD score (approximately 10.5) among the same set of firms.

The D-SOCIAL-KLD measure is able to make a distinction between these firms because it recognizes that Ford is engaging in relatively difficult CSR policies, while Saul Centers is not engaging in easy opportunities to correct socially irresponsible actions—and vice versa. To assess whether or not the differences between our D-SOCIAL-KLD scores better reflect CSR realities than the KLD Index, we examine how our results compare with existing critiques of the KLD Index.

We focus on research that makes specific claims about whether or not certain firms were treated too harshly or too generously in the construction of the equally weighted KLD Index. Those claims come from Entine (2003), whose critiques are broad-ranging, and Delmas and Blass (2010), whose critiques focus primarily on environmental manifestations of CSR. Both of these articles claim that the KLD Index treats some firms too generously and others too harshly, naming some firms explicitly or otherwise making claims about industries as a whole. To assess the validity of our IRT-based measurement model, we can compare these authors' claims about certain firms' performance in the KLD Index to their performance in the D-SOCIAL-KLD scores.

Specifically, Entine (2003) predicted that the KLD Index rated firms in the technology industry too generously given the secretive nature of their businesses. The left panel of Figure 5 focuses on 17 technology firms in the S&P 500, including Apple, and depicts scatterplots of the relative rankings of these firms' firm-year observations on our D-SOCIAL-KLD score versus those on the KLD Index from 1991 through 2003 (the year Entine published his paper). The 45-degree line indicates where firm-year observations would fall if there were no differences between the KLD Index values and those in our D-SOCIAL-KLD scores. Sixty-eight percent of our D-SOCIAL-KLD score predictions are consistent with Entine's claim about technology firms. Entine (2003) also makes a number of other predictions about firms that are rated both too generously and too harshly in the KLD Index; our D-SOCIAL-KLD scores generally match his predictions.

Entine (2003) and Delmas and Blass (2010) both suggest that it is particularly hard for the KLD Index to rate firms in industries where the opportunities

DOI: 10.1002/smj

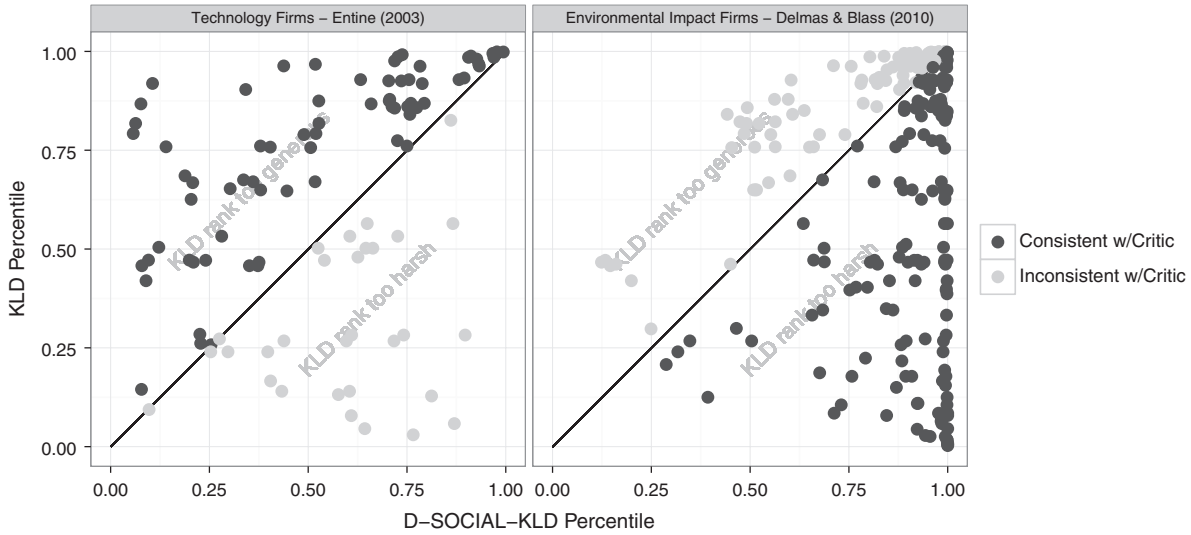

Figure 5. Relative rankings, D-SOCIAL-KLD score versus KLD Index

to avoid environmental degradation are rare, but where the positives are difficult to observe. In an analysis of 15 firms in the chemical sector, Delmas and Blass (2010) make individual cases for why some firms are treated too harshly while others are treated too generously. The right panel of Figure 5 illustrates the analysis for this set of firms for the years 1991–2010 (2010 being the year Delmas and Blass published their study), using the same approach as in the left panel. Our measure agrees 63 percent of the time with their assessment that these firms may have frequently been rated too harshly, given structural issues that make them polluters with problems that are difficult to solve.

The direct comparisons between the KLD Index and the D-SOCIAL-KLD score highlight two final important points about IRT models. First, *relative to the additive KLD Index, a Bayesian IRT analysis offers a much more nuanced (and different) picture of firms*, especially for firms with a large number potentially "offsetting" strengths and concerns, and that cluster around the modal zero value. Second, *because it does not treat every underlying CSR indicator equally, the IRT-based D-SOCIAL-KLD score reflects a number of limitations in the KLD Index previously identified by critics*.

Predictive capabilities of D-SOCIAL-KLD scores

A tougher test of superior validity is a horserace to see whether D-SOCIAL-KLD scores do a better job than the KLD Index itself of predicting behavior on new indicators for CSR "strengths" or CSR "concerns" that are added to the KLD Index as additional components. We can make this test even "tougher" by adding a factor-analysis-derived score to the horserace as well. The year 2010 is a particularly fertile one in which to conduct such a horserace, since KLD added seven new indicators to its database that year across different categories: governance structures (CGOV_CON_K); community engagement (COM_STR_H), employment of underrepresented groups (DIV_STR_H), environmental impact of products and services (ENV_CON_G), biodiversity and land use (ENV_CON_H), operational waste (ENV_CON_I), and operations in Sudan (HUM_CON_H). We can use these new indicators to compare the performance of the 2009 KLD Index score, a D-SOCIAL-KLD score constructed from a Bayesian IRT routine run on KLD data through 2009, and a score based on a one-dimensional factor analysis of the 2009 KLD data (as factor analysis on the entire dataset through 2009 is computationally infeasible). By construction, none of these scores incorporate information about the *ex post* addition of the new indicators.

With these scores, we can run three bivariate probit regression models with each of the new indicators as dependent variables. There is one lone explanatory variable in the three models run for each indicator: the 2009 D-SOCIAL-KLD score,

DOI: 10.1002/smj

the 2009 KLD Index score, or the 2009 factor analysis score.

To assess how well the competing measures perform, we study the fundamental tension between a predictor's "sensitivity," or true positive rate, as a function of its "fall-out," or false positive rate. Suppose that a categorization scheme predicted a "1" whenever a probit model's predicted probability was above some threshold c, which can fall anywhere between 0 and 1. For example, if $c = 0.25$, a predicted probability of 0.15 would be assigned a 0, while a predicted probability of 0.4 would be assigned a 1. There are four possible outcomes: a predicted 0 and a true 0 (a "true negative"), a predicted 1 and a true 0 (a "false positive"), a predicted 0 and a true 1 (a "false negative"), and a predicted 1 and a true 1 (a "true positive"). A good predictor is one with many true positives and true negatives, but very few false positives and false negatives. Of course, these results are a function of the selected c, and the c that yields few false positives (that is, a conservative c) is also one that will generate many false negatives. Accordingly, it is important to consider how these results turn out across *all* possible selections of c. This is just the sort of analysis we conduct.

We summarize these results in Figure 6 using a graphical tool known as a receiver operating characteristic (ROC) curve (Krzanowski and Hand, 2009). An ROC curve captures how well a predictor does in a binary classification system by plotting the predictor's "sensitivity," or true positive rate, as a function of its "fall-out," or false positive rate, for varying levels of criterion value c. A good predictor has high sensitivity even at low levels of fall-out. Graphically, a good ROC curve is one that tends to the northwestern corner of the graph as it moves from west to east. We include a 45-degree line as a point of comparison; this line represents the baseline of random guessing as a predictor, such that being as far as possible to the north and west of it as you move up the line is more desirable. Therefore, a "good" predictor is one that has a large area under the curve, and we can use these areas to compare the relative predictive power of the three metrics.

It is clear from Figure 6 that D-SOCIAL-KLD wins the battle handily over factor analysis, with the KLD Index being a clear loser. D-SOCIAL-KLD has the highest area under the curve in six of the seven horseraces, with factor analysis winning in the final race predicting human rights violations by way of operations in Sudan (HUM_CON_K). But, this is perhaps the least meaningful (methodologically) of the seven indicators, as there is very little variation in the dependent variable (as indicated by the very abrupt nature of the ROC curve). In cases with the most variation in the dependent variable—employment of underrepresented groups (DIV_STR_H), concerns about the environmental impact of products and services (ENV_CON_G), and concerns about biodiversity and land use (ENV_CON_H)—D-SOCIAL-KLD is the best predictor. D-SOCIAL-KLD wins its biggest victory over both factor analysis and the KLD Index in predicting corporate governance structures (CGOV_CON_K)—an indicator with moderate variability.

How can we explain the marked advantage of both the D-SOCIAL-KLD score and factor analysis over the KLD Index in 2010? The answer is that by 2010, the firms in the dataset have become more heterogeneous thanks to the addition of firms outside the S&P and Domini indices after 2001; both IRT and factor analysis can more easily handle this sort of heterogeneity. To see why, note that both IRT and factor analysis assume a continuum of underlying corporate responsibility. As we add more and more heterogeneous firms to the data, factor analysis and IRT can better estimate the underlying structure of responsibility because they have more information to work with.

The KLD Index, on the other hand, has no such advantage, since each firm's score is calculated independently of all other firms. Moreover, the equal weighting assumption built into the KLD Index is even more tenuous as new indicators are added. The extent of KLD Index underperformance is still somewhat surprising; in several of the horseraces, the ROC curve for the KLD Index dips below the 45-degree line, indicating that the KLD Index does *worse* than random guessing.

What about the D-SOCIAL-KLD's defeat of factor analysis? This is due to IRT being able to incorporate information from the entire range of data through 2009, thereby allowing the model to "learn" from the over-time data in a way that factor analysis typically cannot. Specifically, the model takes into account past behavior, and more importantly, *trends in that behavior*.

DOI: 10.1002/smj

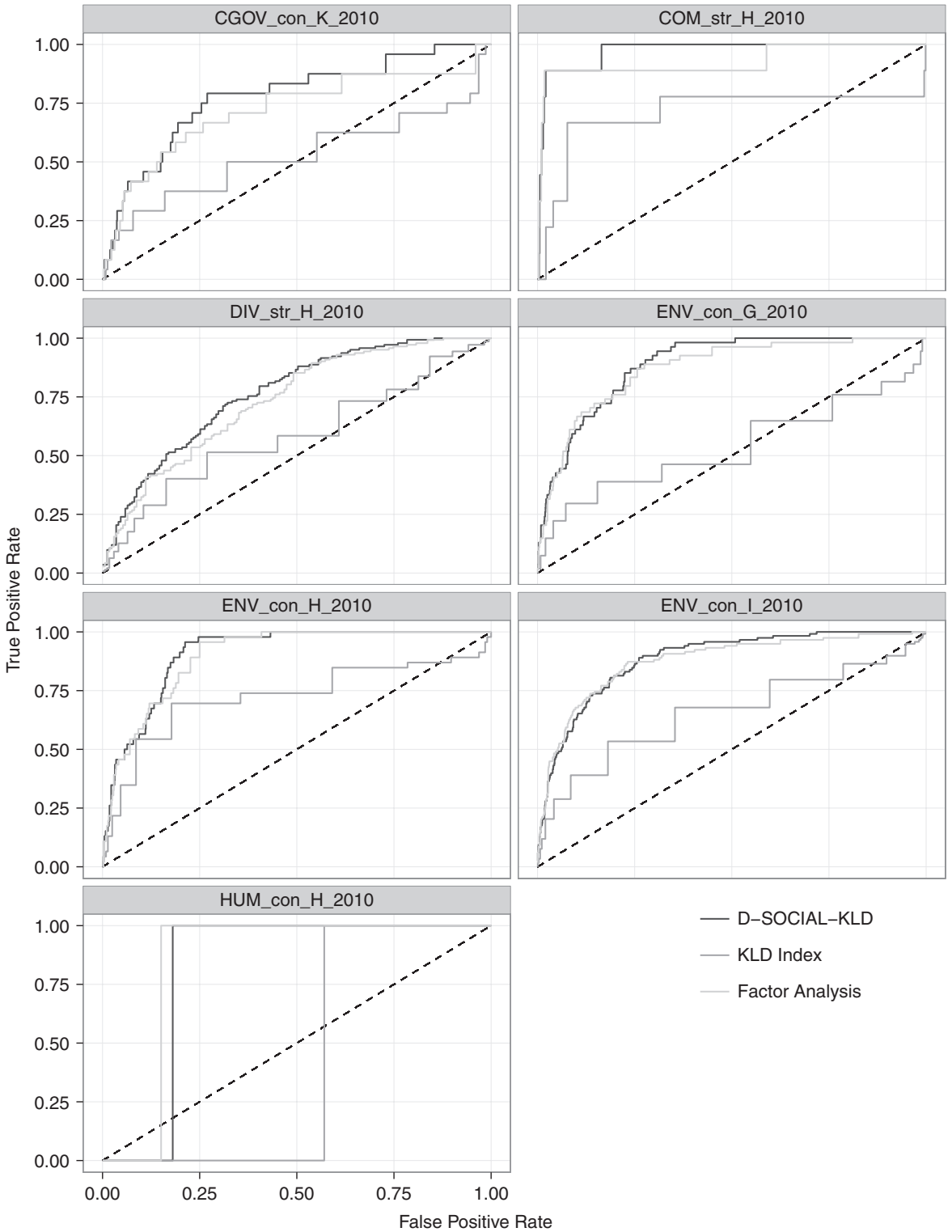

Figure 6. ROC curves summarizing relative predictive power of CSR measures

DOI: 10.1002/smj

CONCLUSION

In this article, we have demonstrated the usefulness of item response theory modeling for strategic management researchers by applying it to commonly used corporate social responsibility data. Of course, not all readers of *SMJ* intend to work in the area of CSR or CSP, but this article's reach extends much farther: our study is useful not only for researchers who want to use our improved CSR/CSP data for their own work or to revisit existing work, but also for researchers who want to adapt the IRT model for new applications. IRT models take full advantage of the data available to the researcher. They produce better measures of constructs than simple additive indices utilizing the same underlying data—and better measures than other data reduction techniques such as factor analysis. Furthermore, they provide a better sense of how reliable the measures they generate are.

Focusing on the data we analyzed in this article, our method shows that the existing additive indices using KLD data sometimes overstate a firm's CSR/CSP levels, and sometimes understate it, often in unexpected ways. Our analysis also shows that some firms are easier to distinguish on CSR/CSP grounds than others, a fact that is lost when looking at additive indices that do not account for measurement error. We also show that the D-SOCIAL-KLD scores produce a more nuanced measure of CSR/CSP than the KLD Index. This is most vividly demonstrated by looking at the big differences in D-SOCIAL-KLD scores for firms with identical KLD Index scores of 0.

Our article contributes to the CSR and CSP literatures in three ways. First, the data we generated in this article opens up new avenues of inquiry in addition to opportunities to revisit earlier work. In the article, we show how our basic model can assess previous critiques about the KLD Index measure—that it is too generous to some firms and too harsh with others—and speak to ongoing debates in the literature on CSR/CSP measurement itself.

Second, because the IRT framework is flexible, it can incorporate additional information into the statistical estimation of CSR and CSP measures. As Chatterji *et al.* (Forthcoming) show, there are other measures and datasets on CSR/CSP beyond the commonly used KLD STATS data highlighted in this article, and the aggregate measures in each often diverge, raising questions about whether they are measuring the same construct. IRT models could incorporate data from these other datasets or help make better comparisons between competing measures. Application-specific measures of IRT modeling for CSR- and CSP-related topics also become readily implementable due to this flexibility. For instance, researchers interested in learning whether environmental regulations influence levels of CSR or CSP could incorporate these rules into the model. An analyst may also want to determine whether there is more than one "dimension" to CSR or CSP. Perhaps environmental issues reflect a different sort of CSR than how workers are treated (e.g., Mattingly and Berman, 2006).

Likewise, some researchers may be interested in using IRT-based methods to further explore whether or not actions that potentially inflict social harm represent a different dimension than actions that potentially provide a social benefit. Mattingly and Berman (2006) explored this question with factor analytic methods in an attempt to resolve a debate about whether or not imposing a single dimension on corporate social action (CSA) as a construct is empirically valid. Baron, Harjoto, and Jo (2011) use only KLD strengths to measure CSP, as they argue that weaknesses are tapping another construct (social pressure). Future work can explore these issues using the IRT approach.

Third, in this article, we focused on the firm-level scores that come out of the IRT model, but in future work, we plan to look at the underlying items themselves. Which are "easy"? Which are "hard"? Do firms appear to adopt these items strategically based on these differences? For instance, Matten and Moon (2008) theorize about how differences between what they call implicit CSR (akin to what our D-SOCIAL-KLD score measures) and what they call explicit CSR (akin to what the KLD Index measures) could be important drivers of strategic action. More generally, our model is sufficiently broad to be able to incorporate simultaneously any number of diverse motivations for individual firms' CSR-related practices found in the literature, including, but not limited to, moral or values-based motivations (e.g., Bansal, 2003); mimetic motivations (e.g., DiMaggio and Powell, 1983; Matten and Moon, 2008); legitimacy concerns (e.g., Bansal and Roth, 2000); managerial-agency-based motivations (e.g., Hemingway and Maclagan, 2004; Hong and Minor, 2013); institutional motivations (e.g., Campbell, 2007; Hoffman, 1999); responsiveness to activists (e.g., Bansal and Roth, 2000; Baron,

DOI: 10.1002/smj

2001; Baron and Diermeier, 2007; Eesley and Lenox, 2006; Lyon and Maxwell, 2011; Reid and Toffel, 2009); insurance-based motivations (e.g., Godfrey, 2005; Godfrey, Merrill, and Hansen, 2009; Minor, 2013; Minor and Morgan, 2011); and strategic or instrumental motivations (e.g., Bansal, 2003; Bansal and Roth, 2000; Kim and Lyon, 2011; Lyon and Maxwell, 2011).

Our article also has implications for researchers seeking to create *new* measures of management or strategy phenomenoa. Though there will always be disagreement about which items should and should not be part of a measure's construction, the IRT model can help sort out competing claims, rather than forcing the analyst to rely on intuition or guesswork. The result will be more reliable indicators on which important empirical analyses of key phenomena can be built. As we noted at the outset of this article, several key measures in management, including those for corporate governance and entrepreneurial orientation, are constructed based on a set of items or actions that are aggregated to create indices. In the area of corporate governance, for instance, the IRT model could address which parts of the "G-index" should be part of that corporate governance score, and which should not.

It is fair to ask of a computationally intensive measure like IRT, "Is the complexity worth it?" Some scholars value methodological rigor and technical sophistication in and of itself; however, that is not our purpose here. Rather, we propose a new method because measurement matters. Taking data values as given—that is, ignoring the modeling assumptions that underlie our data—can have serious detrimental consequences on our tests of substantive theory (Jacoby, 1999). Our approach yields a measure that provides *more* information than the original data, rather than less. The KLD Index, despite its apparent simplicity, imposes a model on the data—an equal-weighted index. The D-SOCIAL-KLD scores we create impose a different model on the data, and we have shown that it outperforms the KLD Index and other measurement models in many ways.

There are very real consequences for using an inferior measure, including the risk of false positives and false negatives. The D-SOCIAL-KLD scores correlate at only 0.195 with KLD Index values, suggesting that replacing a KLD Index measure with a D-SOCIAL-KLD score as an independent variable in a regression is likely to produce very different substantive implications from coefficient estimates. Moreover, taking seriously the measurement error in our estimates (something that cannot be estimated for the KLD Index) is likely to produce larger standard errors for coefficient estimates.

Substantively, Bayesian Dynamic IRT approaches to CSR/CSP measurement are likely to produce more meaningful empirical results when (1) researchers aim to make over-time comparisons within a given firm, since the various underlying items can be more or less important in different years (which the KLD Index does not take into account); and (2) researchers attempt to make comparisons across different types of firms, since the KLD Index does not take into account that firms in various industries might have a greater advantage in scoring well on underlying KLD items than others. More generally, we believe that IRT-based measures of CSR-related constructs may bring greater clarity to much of the extant literature, including the longstanding corporate social performance-corporate financial performance debate.

Of course, there are cases where IRT's complexity and associated computational intensity may not pay off. For example, if you are working with a relatively homogenous set of firms, with a single year of data, with indicators that you have special reason to believe are roughly equal in weight, or simply are not concerned about measurement error, then factor analysis or maybe even a simple additive index may work well for your purposes. But, given the typical research in strategic management, we suspect that these situations will be the exception rather than the rule. Still, even if IRT will typically be the preferred way to construct measures when working with multiple indicators of a latent variable, there are many variants on IRT models and many different underlying assumptions and input data that could be used, meaning that there is room for improvement in any IRT-based measure—including ours. The relevant question then becomes: When do the marginal "bells-and-whistles" that can be added to the IRT model stop adding value and contribute *only* complexity?

The overall message of this article is that researchers can advance measurement in many areas of management and strategy research by utilizing IRT models. There are some start-up costs to doing so, but the payoff—more reliable measures that permit the analyst to pursue new research avenues and revisit old ones—strikes us as a worthy investment.

DOI: 10.1002/smj

ACKNOWLEDGEMENTS

We would like to thank the editors, especially Ashish Arora and Will Mitchell, and three anonymous reviewers for many helpful suggestions. We are also grateful to David Baron, Georgy Egorov, Dylan Minor, Sanjay Patnaik, Ken Shotts, Garrett Sonnier, Jörg Spenkuch, and Dennis Yao for their comments and conversations about the article. We also benefited from audience comments and suggestions at Northwestern University's Kellogg School of Management's Political Economy Seminar Series, George Washington University, the 2014 Strategic Management Society International Conference, the 2014 Academy of Management Annual Meeting, and at the 14th Annual Strategy and the Business Environment Conference. Any errors are our own.

REFERENCES

Aguilera RV, Desender KA. 2012. Challenges in the measuring of comparative corporate governance: a review of the main indices. *Research Methodology in Strategy and Management* **8**: 289–321.

Aguilera RV, Jackson G. 2003. The cross-national diversity of corporate governance: dimensions and determinants. *Academy of Management Review* **28**(3): 447–465.

Aguinis H, Glavas A. 2012. What we know and don't know about corporate social responsibility: a review and research agenda. *Journal of Management* **38**(4): 932–968.

Amaeshi K, Osuji O, Nnodim P. 2008. Corporate social responsibility in supply chains of global brands: a boundaryless responsibility? Clarifications, exceptions and implications. *Journal of Business Ethics* **81**(1): 223–234.

de Bakker FG, Groenewegen P, den Hond F. 2005. A bibliometric analysis of 30 years of research and theory on corporate social responsibility and corporate social performance. *Business and Society* **44**(3): 283–317.

Bansal P. 2003. From issues to actions: the importance of individual concerns and organizational values in responding to natural environmental issues. *Organization Science* **14**(5): 510–527.

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

Barnett ML. 2007. Stakeholder influence capacity and the variability of financial returns to corporate social responsibility. *Academy of Management Review* **32**(3): 794–816.

Baron DP. 2001. Private politics, corporate social responsibility, and integrated strategy. *Journal of Economics and Management Strategy* **10**(1): 7–45.

Baron DP, Diermeier D. 2007. Strategic activism and nonmarket strategy. *Journal of Economics and Management Strategy* **16**(3): 599–634.

Baron DP, Harjoto MA, Jo H. 2011. The economics and politics of corporate social performance. *Business and Politics* **13**(2)Art. 1.

Bock RD. 1997. A brief history of item response theory. *Educational Measurement: Issues and Practice* **16**(4): 21–33.

Bonardi JP, Holburn GLF, Vanden Bergh RG. 2006. Nonmarket strategy performance: evidence from U.S. electric utilities. *Academy of Management Journal* **49**(6): 1209–1228.

Boyd BK, Gove S, Hitt MA. 2005. Construct measurement in strategic management research: illusion or reality? *Strategic Management Journal* **26**(3): 239–257.

Campbell JL. 2007. Why would corporations behave in socially responsible ways? An institutional theory of corporate social responsibility. *Academy of Management Review* **32**(3): 946–967.

Carroll AB. 1979. A three-dimensional conceptual model of corporate performance. *Academy of Management Review* **4**(4): 497–505.

Carroll AB. 1999. Corporate social responsibility: evolution of a definitional construct. *Business and Society* **38**(3): 268–295.

Carroll AB. 2009. A history of corporate social responsibility: concepts and practices. In *The Oxford Handbook of Corporate Social Responsibility*, Crane A, McWilliams A, Matten D, Moon J, Siegel D (eds). Oxford University Press: Oxford, UK, online edition.

Chatterji A, Durand R, Levine D, Touboul S. Forthcoming. Do ratings of firms converge? Implications for managers, investors, and strategy researchers. *Strategic Management Journal*.

Chatterji AK, Levine DI, Toffel MW. 2009. How well do social ratings actually measure corporate social responsibility? *Journal of Economics and Management Strategy* **18**(1): 125–169.

Christensen J, Murphy R. 2004. The social irresponsibility of corporate tax avoidance. *Development* **7**(3): 37–44.

Clinton J, Jackman S, Rivers D. 2004. The statistical analysis of roll call data. *American Political Science Review* **98**(2): 355–370.

Covin JG, Slevin DP. 1991. A conceptual model of entrepreneurship as firm behavior. *Entrepreneurship: Theory and Practice* **16**(1): 7–24.

Dahlsrud A. 2008. How corporate social responsibility is defined: an analysis of 37 definitions. *Corporate Social Responsibility and Environmental Management* **15**(1): 1–13.

Daily CM, Dalton DR, Cannella AA, Jr. 2003. Corporate governance: decades of dialogue and data. *Academy of Management Review* **28**(3): 371–382.

Das J, Hammer JS. 2005. Which doctor? Combining vignettes and item response to measure clinical competence. *Journal of Development Economics* **78**(2): 348–383.

Delmas M, Blass VD. 2010. Measuring corporate environmental performance: the trade-offs of sustainability ratings. *Business Strategy and the Environment* **19**(4): 245–260.

Delmas M, Etzion D, Nairn-Birch N. 2013. Triangulating environmental performance: what do corporate social responsibility ratings really capture? *Academy of Management Perspectives* **27**(3): 255–267.

Diermeier D. 2011. *Reputation Rules: Strategies for Building Your Company's Most Valuable Asset*. McGraw-Hill: New York, NY.

DiMaggio PJ, Powell WW. 1983. The iron cage revisited: institutional isomorphism and collective rationality in organizational fields. *American Sociological Review* **48**(2): 147–160.

Dowling GR. 2014. The curious case of corporate tax avoidance: is it socially irresponsible? *Journal of Business Ethics* **124**(1): 173–184.

Duhigg C, Barboza D. 2012. In China, human costs are built into an iPad. *New York Times* 25 January.

Duhigg C, Kocieniewski D. 2012. How apple sidesteps billions in taxes. *New York Times* 29 April.

Eesley C, Lenox MJ. 2006. Firm responses to secondary stakeholder action. *Strategic Management Journal* **27**(8): 765–781.

Entine J. 2003. The myth of social investing: a critique of its practice and consequences for corporate social performance research. *Organization and Environment* **16**(3): 352–368.

Faye O, Baschieri A, Falkingham J, Muindi K. 2011. Hunger and food insecurity in Nairobi's slums: an assessment using IRT models. *Journal of Urban Health* **88**(Suppl. 2): S235–S255.

Fox JP. 2010. *Bayesian Item Response Modeling: Theory and Applications*. Springer: New York, NY.

Frederick WC. 1994. From CSR1 to CSR2: The Maturing of Business-and-Society Thought. *Business and Society* **33**(2): 150–164.

Gelman A, Carlin JB, Stern HS, Rubin DB. 2003. *Bayesian Data Analysis* (2nd edn). Chapman & Hall: Boca Raton, FL.

Godfrey PC. 2005. The relationship between corporate philanthropy and shareholder wealth: a risk management perspective. *Academy of Management Review* **30**(4): 777–798.

Godfrey PC, Hill CWL. 1995. The problem of unobservables in strategic management research. *Strategic Management Journal* **16**: 519–533.

Godfrey PC, Merrill CB, Hansen JM. 2009. The relationship between corporate social responsibility and shareholder value: an empirical test of the risk management hypothesis. *Strategic Management Journal* **30**(4): 425–445.

Graafland JJ, Eijffinger SCW, Smid H. 2004. Benchmarking of corporate social responsibility: methodological problems and robustness. *Journal of Business Ethics* **53**(1/2): 137–152.

Griffin JJ, Mahon JF. 1997. The corporate social performance and corporate financial performance debate: twenty-five years of incomparable research. *Business and Society* **36**(1): 5–31.

Hays RD, Lipscomb J. 2007. Next steps for use of item response theory in the assessment of health outcomes. *Quality of Life Research* **16**: 195–199.

Hedeker D, Mermelstein RJ, Flay BR. 2006. Application of item response theory models for intensive longitudinal data. In *Models for Intensive Longitudinal Data*, Walls TA, Schafer JL (eds). Oxford University Press: Oxford, UK; 84–108.

Hemingway CA, Maclagan PW. 2004. Managers' personal values as drivers of corporate social responsibility. *Journal of Business Ethics* **50**(1): 33–44.

Hoetker G. 2007. The use of logit and probit models in strategic management research: critical issues. *Strategic Management Journal* **28**: 331–343.

Hoffman AJ. 1999. Institutional evolution and change: environmentalism and the U.S. chemical industry. *Academy of Management Journal* **42**(4): 351–371.

Hong H, Kubik JD, Scheinkman JA. 2012. Financial constraints on corporate goodness. Working paper 18476, National Bureau of Eeconomic Research, Cambridge, MA.

Hong B, Minor D. 2013. Good (bad) company or good (bad) manager? Exploring the antecedents of CSR. Working paper, Northwestern University, Evanston, IL.

Høyland B, Moene K, Willumsen F. 2012. The tyranny of international index rankings. *Journal of Development Economics* **97**(1): 1–14.

Jackman S. 2000. Estimation and inference are missing data problems: unifying social science statistics via Bayesian simulation. *Political Analysis* **8**(4): 307–332.

Jacoby WS. 1999. Levels of measurement and political research: an optimistic view. *American Journal of Political Science* **43**(1): 271–301.

Johnson VE, Albert JH. 1999. *Ordinal Data Modeling*. Springer-Verlag: New York, NY.

Kim EH, Lyon TP. 2011. Strategic environmental disclosure: evidence from the DOE's voluntary greenhouse gas registry. *Journal of Environmental Economics and Management* **61**(3): 311–326.

Kitzmueller M, Shimshack J. 2012. Economic perspectives on corporate social responsibility. *Journal of Economic Literature* **50**(1): 51–84.

Kotchen M, Moon JJ. 2012. Corporate social responsibility for irresponsibility. *B.E Journal of Economic Analysis and Policy* (Contributions) **12**(1): 55.

Krzanowski WJ, Hand DJ. 2009. *ROC Curves for Continuous Data*. CRC Press: Boca Raton, FL.

Londregan J. 2000. *Legislative Institutions and Ideology in Chile*. Cambridge University Press: New York, NY.

Lord FM, Novick MR, Birnbaum A. 1968. *Statistical Theories of Mental Test Scores*. Addison-Wesley: Reading, MA.

Lumpkin GT, Dess GG. 1996. Clarifying the entrepreneurial orientation construct and linking it to performance. *Academy of Management Review* **21**(1): 135–172.

Lyon DW, Lumpkin GT, Dess GG. 2000. Enhancing entrepreneurial orientation research: operationalizing and measuring a key strategic decision making process. *Journal of Management* **26**(5): 1055–1085.

Lyon TP, Maxwell JW. 2011. Greenwash: corporate environmental disclosure under threat of audit. *Journal of Economics and Management Strategy* **20**(1): 3–41.

Margolis JD, Elfenbein HA, Walsh JP. 2009. Does it pay to be good? A meta-analysis and redirection of

research on the performance between corporate social and financial performance. Working paper, available at: SSRN at http://ssrn.com/absract=1866371 (accessed 20 November 2015).
Margolis JD, Walsh JP. 2003. Misery loves companies: rethinking social initiatives by business. *Administrative Science Quarterly* **48**(2): 268–305.
Martin AD, Quinn KM. 2002. Dynamic ideal point estimation via markov chain monte carlo for the U.S. Supreme Court. *Political Analysis* **10**(2): 134–153.
Matten D, Moon J. 2008. Implicit and explicit CSR: a conceptual framework for a comparative understanding of corporate social responsibility. *Academy of Management Review* **33**(2): 404–424.
Mattingly JE, Berman SL. 2006. Measurement of corporate social action: discovering taxonomy in Kinder Lydenberg Domini ratings data. *Business and Society* **45**(1): 20–46.
Minor D. 2013. The value of corporate citizenship: protection Working paper, Northwestern University, Evanston, IL.
Minor D, Morgan J. 2011. CSR as reputation insurance *Primum Non Nocere*. *California Management Review* **53**(3): 40–59.
MSCI ESG Research. 2012. MSCI ESG States: user guide & esg ratings definition. June.
Muller A, Kräussl R. 2011. Doing good deeds in times of need: a strategic perspective on corporate disaster donations. *Strategic Management Journal* **32**(9): 911–929.
Orlitzky M, Schmidt FL, Rynes SL. 2003. Corporate social and financial performance: a meta-analysis. *Organization Studies* **24**(3): 403–441.
Orlitzky M, Siegel DS, Waldman DA. 2011. Strategic corporate social responsibility and environmental sustainability. *Business and Society* **50**(1): 6–27.
Poole KT, Rosenthal H. 1991. Patterns of congressional voting. *American Journal of Political Science* **35**(1): 228–278.
Rasch G. 1960. *Probabilistic Models for Some Intelligence Tests and Attainment Tests*. Danish Institute for Educational Research: Copenhagen, Denmark.
Reid EM, Toffel MW. 2009. Responding to public and private politics: corporate disclosure of climate change strategies. *Strategic Management Journal* **30**(11): 1157–1178.
Reise SP, Waller NG. 2009. Item response theory and clinical measurement. *Annual Review of Clinical Psychology* **5**: 27–48.
Rowley T, Berman S. 2000. A brand new brand of corporate social performance. *Business and Society* **39**(4): 397–418.
Sharfman M. 1996. The construct validity of the Kinder, Lydenberg & Domini social performance ratings data. *Journal of Business Ethics* **15**(3): 287–296.
Sharfman MP, Fernando CS. 2008. Environmental risk management and the cost of capital. *Strategic Management Journal* **29**: 569–592.
Shleifer A, Vishny RW. 1997. A survey of corporate governance. *Journal of Finance* **52**(2): 737–783.
Strike VM, Gao J, Bansal P. 2006. Being good while being bad: social responsibility and the international diversification of US firms. *Journal of International Business Studies* **37**: 850–862.
Surroca J, Tribó JA, Waddock S. 2010. Corporate responsibility and financial performance: the role of intangible resources. *Strategic Management Journal* **31**(5): 463–490.
Thurstone LL. 1925. A method of scaling psychological and educational tests. *Journal of Educational Psychology* **16**(7): 433–451.
Vanden Bergh RG, Holburn GLF. 2007. Targeting corporate political strategy: theory and evidence from the U.S. accounting industry. *Business and Politics* **9**(2): 1–31.
Votaw D. 1972. Genius becomes rare: a comment on the doctrine of social responsibility pt. 1. *California Management Review* **15**(2): 25–31.
Waddock SA. 2003. Myths and realities of social investing. *Organization Environment* **16**(3): 369–380.
Waddock SA, Graves SB. 1997. The corporate social performance-financial performance link. *Strategic Management Journal* **18**(4): 303–319.
Walls JL, Phan PH, Berrone P. 2011. Measuring environmental strategy: construct development, reliability and validity. *Business and Society* **50**(1): 71–115.
Wood DJ. 1991. Corporate social performance revisited. *Academy of Management Review* **16**(4): 691–718.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Appendix S1. Using item response theory to improve measurement in strategic management research: an application to corporate social responsibility

DOI: 10.1002/smj