Using Item Response Theory to Improve Measurement in Strategic Management Research: An Application to Corporate Social Responsibility

Robert J. Carroll* David M. Primo† Brian Kelleher Richter‡

March 25, 2014

Abstract

We introduce item response theory (IRT) to management and strategy research. IRT explicitly models firms' and individuals' observable actions in order to measure unobserved, latent characteristics. IRT models have helped researchers improve their measures in numerous disciplines. To demonstrate their potential in strategic management, we show how the method improves upon the *de facto* best measure of corporate social responsibility (CSR), the KLD Index, by creating IRT Responsibility scores from the underlying data along with estimates of their accuracy. We show, for instance, that firms like Apple may not be as socially responsible as previously thought, while firms like Walmart may be more responsible than typically believed. We also show that IRT Responsibility scores are better at predicting new CSR activity than the KLD Index.
Keywords: Research Methods, Measurement, Item Response Theory, Bayesian Estimation, Corporate Social Responsibility

*Department of Political Science, University of Rochester, Rochester, NY 14627. rcarroll@mail.rochester.edu
†Department of Political Science, University of Rochester, Rochester, NY 14627. david.primo@rochester.edu
‡McCombs School of Business, University of Texas, Austin, TX 78712. brian.richter@mccombs.utexas.edu

INTRODUCTION

Obtaining and relying upon valid measures is no less fundamental to high quality research in management and strategy than it is in the physical sciences (Venkatraman and Grant, 1986)—which is why poor measurement represents "one of the most serious threats to strategic management research" today (Boyd, Gove, and Hitt, 2005a). 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., Shleifer and Vishny, 1997; Aguilera and Jackson, 2003; Daily, Dalton, and Cannella, 2003), how entrepreneurial (e.g., Covin and Slevin, 1991; Lumpkin and Dess, 1996), or how socially responsible (e.g., Carroll, 1979) a given firm really is? These attributes all represent complicated constructs that may or may not generate observable actions. Moreover, any entrepreneurial act, governance act, or social responsibility act we observe is distinct from the underlying latent attribute we want to measure, as it is only representative of a given firm's performance conditional on its environment drawing out the behavior. Hence, using observables as proxies for latent attributes typically leads to measurement with systematic but unknown error.

This paper aims to answer longstanding questions about how we can improve the reliability and accuracy of such measures. It does so by introducing to strategic management research the tool of item response theory (IRT) modeling, which can be applied to data with multiple observables obtained through surveys, observation, or other means. Then, in an application, it demonstrates how IRT models can improve upon the most widely adopted existing measure of firms' latent corporate social responsibility (CSR). An appendix includes technical details to help future researchers adapt the tool to improve measurement in other strategic management contexts.[1]

[1] To further help other researchers learn how to apply IRT models to improve measurement in other strategic management contexts, the authors plan on making replication software (and underlying code written in R) publicly available on the Internet. To aide researchers interested in applying this improved CSR measure in future CSR studies or in replication studies of past work, the authors also plan on making that data publicly available on the Internet. URLs will be made available upon publication.

IMPROVING UPON "STATE OF THE ART" MEASUREMENT IN STRATEGIC MANAGEMENT RESEARCH

Boyd, Gove, and Hitt (2005a), in a survey of the types of measures used in empirical strategic management studies, argue persuasively that indices and scales provide more reliable measures of latent constructs than any single proxy because they combine information from multiple observables, thereby reducing measurement error from any one noisy signal. Indices, which they define as taking an additive approach to combine information on observable attributes, are slightly more commonly used than scales, which they define as taking a data reduction approach such as factor analysis or principal components analysis (Boyd, Gove, and Hitt, 2005a).

While there are many examples of additive indices used in the strategic management literature, a prominent example is the "G-index," which is used to measure the quality of corporate governance (Aguilera and Desender, 2012). To construct the "G-index" for a given firm, researchers add one point for each of twenty-four observable provisions restricting shareholder rights, such as whether a given firm limits the ability of directors to call special meetings or maintains "golden parachute" severance agreements for top executives (Gompers, Ishii, and Metrick, 2003). The implicit assumption underlying the construction of additive indices used in strategic management, including the "G-index," is that the presence (or absence) of each observable action is an equally good proxy of the underlying attribute we hope to measure. This, of course, is a strong assumption which is difficult to justify theoretically. So, for any given additive index, critics tend to argue that certain observable actions (i) should not be included in the index at all (e.g., Sonnenfeld, 2004), (ii) should be given lower weights in the index (e.g., Bebchuk, Cohen, and Ferrell, 2009), or (iii) should be included in the index only in certain contexts (e.g., Bhagat, Bolton, and Romano, 2008).

Scales using data reduction approaches such as factor analysis make a different set of assumptions than additive indices about the relevance of observable actions vis-á-vis the underlying latent characteristics they attempt to measure. While there are many examples of factor analytic scales used in the strategic management literature, a prominent example is the firm-level "entrepreneurial orienta-

tion (EO) scale" (Lyon, Lumpkin, and Dess, 2000). To construct the EO scale, researchers first survey managers about aspects of their strategic posture including innovativeness, proactiveness, and risk taking; then, the researchers weight responses as a linear combination according to loadings obtained by factor analyzing the data (Covin and Slevin, 1989), possibly ignoring some responses altogether if the factor loadings are too low (e.g., Lumpkin and Dess, 2001; Richard et al., 2004). While an improvement over additive indices, factor analysis is not as flexible as the approach we introduce next.

IRT MODELS

Item response theory (IRT) models can improve upon "state of the art" measurement techniques by generating measures of latent characteristics based upon a richer, theory-driven understanding of how these characteristics are reflected in proxies. Precursors to IRT models were initially developed by Thurstone (1925) in the education literature when he 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).[2] IRT models simultaneously assess both the test questions and the test takers. The data inputted into an IRT model for estimation of latent traits may be responses to a set of questions or a set of other observed measures, such as whether various behaviors occurred or did not occur. The basic two-parameter model for binary (e.g., yes-no; absent-present; 0-1; correct-incorrect) data takes the following form: $\Pr\left(y_{i,j} = 1 \mid \rho_i, \alpha_j, \beta_j\right) = F\left(-\alpha_j + \beta_j \rho_i\right)$.[3]

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 we hope 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

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

[3] IRT models can also accommodate ordinal responses (e.g., a rating on a scale of 1 to 5). Additional parameters can be added to IRT models to handle the specific needs of researchers. For instance, in the analysis of a test, a parameter accounting for guessing might be useful.

each item is (α_j) and how well each item discriminates among individuals (β_j). Using a test analogy, α_j addresses the question "How likely is any respondent to get question j correct?" β_j addresses the question "How well does question j help distinguish between high and low ability levels?"; in other words, do individuals with high ability and low ability (i.e., high and low ρ_is) differ in the probability they will get a question correct?[4]

IRT models have deep roots in psychology (Rasch, 1960; Lord and Novick, 1968; Reise and Waller, 2009) and have made inroads into disciplines including economics (e.g., Høyland, Moene, and Willumsen, 2012) and medicine and public health (Das and Hammer, 2004; Hedeker, Mermelstein, and Flay, 2006; Hays and Lipscomb, 2007; Faye, Baschieri, Falkingham, and Muindi, 2011). The closest analogue to the IRT analysis in this paper, 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 point") on a left-right continuum. Proxies for this latent trait, like party affiliation (e.g., Bonardi, Holburn, and Vanden Bergh, 2006; Vanden Bergh and Holburn, 2007) or relative campaign contributions (e.g., Burris, 2001; Chin, Hambrick, and Trevino, 2013), are often crude (e.g., party is dichotomous for the U.S.) or very indirect (e.g., campaign contributions reflect many other factors besides ideology (Fremeth, Richter, and Schaufele, 2013)). On the other hand, votes on specific bills reflect the revealed preferences of individual legislators. Returning to the test analogy, "questions" in this context are votes on legislation and "answers" are "yea"s and "nay"s.

Just as IRT models use the pattern of answers to test questions to measure student performance, IRT models use the pattern of votes on bills to estimate the ideology of legislators (Poole and Rosenthal, 1991; Jackman, 2000; Londregan, 2000; Clinton, Jackman, and Rivers, 2004). The theoretical foundation for estimation of ideal points is typically a spatial model in which legislators are placed in this space on a left-right continuum based on their ideology. Legislation can also be placed in this

[4]IRT models bear more than a passing resemblance to a common method in strategic management research: factor analysis. In fact, the two approaches are closely related (Takane and De Leeuw, 1987; Kamata and Bauer, 2008). Jackman (2001), referencing the work of Bock, Gibbons, and Muraki (1988), points out that certain forms of item response theory are referred to as "full information item factor analysis." Quinn (2004) develops a measurement model that combines factor analysis and item response theory. That said, there are important differences, including that Bayesian item response modeling can more readily estimate all parameters of interest and can incorporate theory into the statistical estimation.

space, as can current policy (the status quo). Legislators are assumed to vote for a bill if it is closer to their ideal point than the status quo, and against it otherwise.

Taking work on legislatures and adapting it to the courts, Martin and Quinn (2002) explicitly apply IRT to a dynamic setting in which ideal points of Supreme Court judges voting on cases are allowed to vary over time. As they note, this is a deviation from the educational testing literature, in which underlying traits like intelligence are assumed to be fixed. The underlying theoretical and statistical foundation for ideal point estimation works equally well in business settings, where instead of voting on legislation, managers or firms can adopt certain business policies or behaviors or not, so we adopt Martin and Quinn's approach in what follows.

IRT measurement models can aid strategic management research by generating not only valid point estimates but also valid confidence intervals around those point estimates, thereby explicitly acknowledging the level of uncertainty in the measure for a given firm or individual. This is of both fundamental and instrumental concern: we care about the quality of our estimates absolutely, but we also want to ensure that those future scholars interested in using our measures on the right-hand side of regression equations are able to explicitly account for the appropriate level of confidence to put in those measures. In addition, IRT models allow the researcher to estimate the weights on items in an index rather than assume that they are all equal. As we will see, this is a very important feature of these models.

Two recent advances in IRT modeling also are notable because they make IRT models more relevant for the strategic management context and explain their recent explosion in popularity as a measurement tool. First, given increased computing power and the introduction of Bayesian methods such as Markov chain monte carlo (MCMC) simulation techniques, estimating the large number of parameter values in IRT models is no longer intractable (Albert, 1992; Martin and Quinn, 2002; Treier and Jackman, 2008). Second, IRT models have been adapted to dynamic contexts such that our measures of latent characteristics can vary within firms and within individuals over time (Martin and Quinn, 2002).

In short, a theory-driven, dynamic IRT approach to measurement in strategic management re-

search contexts allows the analyst, using a basic model, to assess:

- whether differences between individuals and firms in traditional measures are real or due to systematic but unknown measurement error
- how individual firms and groups of firms change over time
- whether, and by how much, items in an index are better/worse at distinguishing among firms.

To demonstrate the power of IRT in a strategic management context, we will adapt the Martin and Quinn (2002) model to improve upon a measure of the corporate social responsibility (CSR) construct. We keep the presentation and application of the IRT model in this paper intentionally simple by estimating a two-parameter Bayesian dynamic model on binary data so that we can best communicate the key features of IRT to a strategic management audience without distraction. We leave alternative, more complicated approaches to modeling IRT-based estimates of CSR to future dedicated applications that could benefit from incorporating additional features.

MEASURING CORPORATE SOCIAL RESPONSIBILITY ACTIVITY

While we could have demonstrated how IRT models can improve upon many other existing indices or scales in strategic management, we believe CSR is an ideal place to start because: (i) there is general agreement on a *de facto* standard measure of CSR activity in the Kinder, Lydenberg, Domini (KLD) index (Waddock, 2003), (ii) despite the widespread use of this *de facto* standard measure, researchers using it openly acknowledge its limitations (e.g., Sharfman and Fernando, 2008; Kacperczyk, 2009), and (iii) numerous published articles have focused explicitly on critiquing the *de facto* standard measure (e.g., Entine, 2003; Chatterji, Levine, and Toffel, 2009; Delmas and Blass, 2010). Hence, we view improving upon the KLD Index measure of CSR activity—in addition to introducing IRT to management and strategy researchers—as an important contribution by itself, since the research community has openly called for a better measure.

Corporate social responsibility (CSR) 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. Our focus here is on corpo-

rate responses to the CSR construct and their measurement in the literature.[5]

The CSR construct is a complicated one that may be manifest in a number of different behaviors depending upon firm-specific factors and competing definitions (Carroll, 1979, 1999; Dahlsrud, 2008). Data-driven CSR work began in the 1980s, if not earlier; however, this early CSR research "was plagued with measurement problems, because few good measures existed for the multidimensional construct," according to Surroca, Tribó, and Waddock (2010), who also note that "researchers tended to select a single item as a proxy." Moreover, a number of empirical researchers, following Frederick (1994), decided to sidestep the CSR construct altogether by limiting interpretations of their findings to "narrower and more technical" definitions they labeled corporate social performance (CSP). Since then, there have been myriad academic debates about the proper definitions of the CSR and CSP terms, although no standard has emerged. Rather, as McWilliams, Siegel, and Wright (2006) note, CSP is now "often used as a synonym for CSR."[6]

Things began to look up for the measurement of corporate responses to the CSR construct when Waddock and Graves (1997) introduced the KLD dataset to academic researchers. The KLD data was 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.[7]

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 to create the commonly used composite measure. The KLD STATS dataset includes over 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"

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

[6]Orlitzky, Siegel, and Waldman (2011b) recently noted that the preponderance of semantic debates has "hampered scientific progress" in understanding the activity itself and drawn the focus away from what the measures used in empirical work actually capture.

[7]See MSCI ESG Research (2012) for details on the creation and construction of the KLD database.

which proxy social responsibility, and other indicators as "concerns" which proxy social irresponsibility. For example, for a diversity strength labeled "Gay & Lesbian Policies," a firm receives a score of 1 if "it provides benefits to domestic partners of its employees," and for a human rights concern labeled "Operations in Burma," a firm receives a 1 if "the company has operations or direct investment in, or sourcing from, Burma. Hence, the structure of the KLD data, and that of most data underlying additive indices in management and strategy research, mirrors that used in other applications of IRT measurement models, such as answers to test questions used to measure students' abilities or, more similarly, votes on bills used to measure legislators' ideologies.

The KLD dataset is "the *de facto* research standard" (Waddock, 2003) in this literature: among the articles published in *SMJ* between 2008 and 2012 that included a numerical measure of a CSR-related construct in their analyses, nearly 85 percent either referenced or used the KLD data. In these papers using the KLD data, an aggregated version of the KLD Index was the most common measure derived from the underlying dataset. An overwhelming majority of those articles nevertheless critiqued some aspect of the KLD Index. Moreover, a large set of articles has emerged where the primary purpose is to critique or assess the validity of KLD, particularly as an additive equal-weight index. Those articles include Sharfman (1996), Griffin and Mahon (1997), Rowley and Berman (2000), Entine (2003), Graafland, Eijffinger, and Smid (2004), Mattingly and Berman (2006), Chatterji, Levine, and Toffel (2009), Delmas and Blass (2010), Walls, Phan, and Berrone (2011), and Delmas, Etzion, and Nairn-Birch (Forthcoming). These critiques tend to be on the same grounds as those for other equally-weighted index measures alluded to in the introduction, echoing the sentiment that "there is a flaw in the assumption that each [KLD] strength category [indicator] and each [KLD] concern category [indicator] are equal" (Sharfman and Fernando, 2008). We will turn back to these critiques after presenting our new CSR measure.

APPLICATION: OUR MODEL AND DATA

In this section, we introduce the key theoretical elements of our IRT model for CSR 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.[8] More precisely, we devise a model focusing on the utility, or benefit, that a firm receives from adopting (or not adopting) a particular CSR policy (e.g., a recycling program).[9]

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.[10] Firm i's utility is a function of its underlying, latent level of CSR ($\rho^d_{i,t}$), the level of CSR 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: $u^d_{i,j,t} = -\left|\rho^d_{i,t} - \tau^d_{j,t}\right|^2 + \xi^d_{i,j,t}$. 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 non-adoption is consistent with the firm's underlying responsibility. It is straightforward to reverse the logic when thinking about 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

[8] Given space limitations and our primary interest in demonstrating how the IRT approach can improve upon measures in management broadly, technical details of how we applied an IRT model to the KLD data and how we estimated parameter values in that model will be provided in an appendix.

[9] The model in this section draws from a model developed by Clinton, Jackman, and Rivers (2004) in the context of legislative voting.

[10] We note that this conception of the utility gained from CSR is sufficiently broad to be able to incorporate simultaneously any number of diverse motivations for individual firms' CSR 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., Hoffman, 1999; Campbell, 2007); responsiveness to activists (e.g., Bansal and Roth, 2000; Baron, 2001; Eesley and Lenox, 2006; Baron and Diermeier, 2007; Reid and Toffel, 2009; Lyon and Maxwell, 2011); insurance-based motivations (e.g., Godfrey, 2005; Godfrey, Merrill, and Hansen, 2009; Minor and Morgan, 2011; Minor, 2013); and strategic or instrumental motivations (e.g., Bansal and Roth, 2000; Bansal, 2003; Kim and Lyon, 2011; Lyon and Maxwell, 2011). While our IRT approach can parse out these and other motivations or underlying qualities of CSR, we leave that task—which requires analysis of the policy-specific parameters in the model below and estimates of the dimensionality of the data—for later work.

$z_{i,j,t} = u^A_{i,j,t} - u^R_{i,j,t}$. We can substitute the formulas above into this equation and simplify as follows:

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

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. Here $\alpha_{j,t}$ is the *difficulty* parameter for adopting policy j in time period t, which can be thought of loosely as the level of effort required for a firm to implement policy j independent of its latent level of CSR. $\beta_{j,t}$ is the *discrimination* parameter for adopting policy j in time period t. If $\beta_{j,t}$ is positive, then more 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 paper.

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" and "nay" votes depending on the matter at hand, so too is a "responsible'' firm one that follows a particular pattern of corporate governance. 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, 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. While we reserve the nuts and bolts of the estimation for the appendix, we discuss the key feature of our approach—Bayesian estimation—in the next section.

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 policies to adopt given some unobservable level of CSR. We adopt a Bayesian mode of inference for both theoretical and pragmatic reasons. Prior to laying out those justifications, however, some brief review of the Bayesian model of inference is warranted.[11]

While the divergence between the two approaches is often overstated, the Bayesian approach is very different from the classical, frequentist approach to inference in this context. A frequentist approach treats the parameters enumerated above as fixed and then estimates their values using the available data. The frequentist parallel most relevant for our purposes is maximum likelihood estimation, which is used for techniques like probit.

The Bayesian approach, on the other hand, 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." This updating makes use of a result from probability theory, Bayes' rule (hence, the term Bayesian estimation), which allows one to "update" about the likelihood of an event based on the observation of new data. Using a technique known as Markov chain monte carlo (MCMC), detailed in the Appendix, the Bayesian analyst can obtain simulated distributions of the parameters of interest and then obtain meaningful results—say, the posterior mean—based on those simulated distributions.

All modeling requires assumptions, and the 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 dictates how closely information from the previous period relates to information in the current period and is estimated as part of the model. If this value is very small, then the time

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

series of IRT Responsibility for a firm over time approaches a constant value. If it is very large, then the time series is essentially unrelated to itself across time. Martin and Quinn (2002) observe that this is a "happy median" between one extreme (not modeling changes over time at all) and the other (not allowing one time period's responsibility to be related to the next); we follow their lead because firm leadership is a relatively "sticky" quality: decision-making structures remain the same for decades, and executives enjoy relatively long tenures.

Importantly, for the difficulty and discrimination terms, we do *not* model dynamic effects in policy-specific attributes. Instead, we treat each observable as a "new case" in each year. To be sure, the analyst could utilize a random walk model like we utilize for the IRT Responsibility score. But, there are practical and theoretical reasons for not doing so. Practically, this already massive computing problem becomes much more complicated by adding dynamics to the items. Theoretically, we want to allow for the most flexibility in the estimation of the difficulty and discrimination parameters, since a key purpose of our enterprise is to learn what CSR is (i.e., improve on existing measures).[12]

Bayesian approaches have several other practical benefits beyond the incorporation of dynamics. First, unlike frequentist approaches used to estimate, for instance, a probit model, MCMC does not require the maximization of a function. In an application like ours, with several thousand parameters, this is a major advantage. It is, in general, far simpler to simulate the posterior distributions of those parameters than it is to attempt to numerically optimize over all of them.

Second, the estimation of simulated distributions means that we can get a more nuanced picture of how accurate our estimates are, compared with more traditional approaches. The existing literature makes it abundantly clear that CSR is imperfectly measured—and Bayesian approaches can help us understand the nature of that uncertainty by quantifying it.

Third, the Bayesian approach can easily handle missing data, which as we note below, is of particular importance in our application and all others that work with firms that may be represented in a dataset in one year but not another (e.g., datasets focused on firms in a particular stock index).

Fourth, given the assumptions about dynamics, the entire routine is implementable using free,

[12] As it happens, these parameter estimates do exhibit stickiness, suggesting a role for dynamics in future work, subject to computing limitations.

standard software in the `MCMCpack` library in the R statistical computing environment.[13] More complicated models (e.g., those assuming that difficulty and discrimination parameters are dynamic) are estimable with some minor coding modifications.

Finally, given the novelty of our application, we do not incorporate prior information into the model save for how the parameters interact with one another; however, future 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 (Statistical Tools for Analyzing Trends in Social and Environmental Performance) data, which provides annual "snap-shots of the environmental, social, and governance performance of companies rated by KLD Research & Analytics, Inc." (KLD Stats, 2008). As discussed above, the KLD data represent the disciplinary standard for corporate social action. We cover the entire breadth of available data, from 1991–2012.

The KLD data include a wide variety of indicators, over 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 1991–2000, 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 U.S., 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,

[13] Specifically, we obtain our results using the `MCMCdynamicIRT1d()` routine. Given the size of the data and the amount of missingness, care must be taken to assign proper starting values, but the analysis is otherwise quite standard.

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. 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 data set. 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. The `MCMCpack` routine, which is a straightforward application of the standard data augmentation techniques proposed by Albert and Chib (1993), treats 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 responsibility 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 (whether 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.

The routine takes an "agnostic'' approach to missing values for indicators when a firm does not have missing values for all other indicators in a given year; these are assumed to arise from an untruncated normal distribution, thus imposing no "sign'' attributed to a 1 or a 0. This allows the routine to "fill in'' the required value in the absence of data while still taking the other, observed data into account. In contrast, observed values are assumed to have underlying utilities that are distributed truncated normal at zero (the positive side for 1s and the negative side for 0s). Importantly, this ensures that missing observations do not influence our estimates of those parameters that are included, as (on average) these draws do not affect estimates of the posterior's underlying parameters. Indeed, the only effect here is to increase the uncertainty of the estimates by adding "noise'' conditioned on the MCMC

routine's estimates of those parameters.

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 of each of the 1,610 α terms and of the 1,610 β terms (a pair for each observed policy) along with the values of each of the 40,505 ρ terms (one for each observed firm in each year). Our final result is a simulation of the complete joint posterior distribution of all 43,725 parameters[14] in the model. 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 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.

APPLICATION: RESULTS

The IRT model takes a very large data matrix full of binary responses and missing observations and produces what we call IRT Responsibility scores linking observations from multiple years. While the overall distribution of IRT Responsibility scores is roughly centered around zero, the zero point itself has no innate meaning.[15] What matters in these scores, just as with KLD Index scores, 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 IRT Responsibility scores we estimated for all firms in 1991 in panel (a) and in 2005 in panel (b) of Figure 1.

[Figure 1 about here.]

We choose to display 1991 in panel (a) 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

[14] We provide 2,500 draws from the joint posterior distribution, meaning that the final data matrix has 109,312,500 unique elements.

[15] In the language of statistics, these are interval data, not ratio data.

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, in panel (a) the bottom-most observation in 1991 corresponds with Golden West Financial, while the highest score goes to DuPont.[16]

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 intervals in frequentist statistics). In 1991, there is substantial overlap in the inter-percentile ranges for many firms, especially in the middle of the pack—as fewer than 30 percent of firms can be said to have a latent level of CSR greater than the median and fewer than 5 percent can be said to have a latent level of CSR lower than the median.[17] This overlap indicates that it is difficult to distinguish between the level of CSR for 65 percent of firms in 1991. This takes us to the first two lessons we glean from the Bayesian estimation of our IRT model:

1. Firm-to-firm comparisons of CSR using the KLD data should proceed with caution unless the differences in any measure are sufficiently large.
2. Researchers should explicitly account for measurement error when incorporating the KLD data 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.

We also point out that in general the results, like in panel (a) for 1991, look something like the cumulative density function (CDF) for a normally distributed variable. This basic pattern holds across all years. For instance in 2005, shown in panel (b), rather than the dots sitting nearly vertically on top of each other as in panel (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

[16] Our finding that DuPont is the company with the highest level of CSR 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–suggesting that our measure is valid. We elaborate on this point later.

[17] The reader will note that the firms toward the top of the graph tend to be simulated with more precision than the firms toward the bottom of the graph. This occurs because many of the firms toward the top have been covered by KLD in more years than those that fall towards the bottom, many of which exit the S&P 500 and Domini indices in the 1990s. This is especially true for relatively smaller financial firms and savings and loan firms like Golden West, which tend to have lower scores and tend to exit the data more often than other kinds of firms.

time allows us to make more 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 panels (a) and (b) of 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 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.[18]. This brings us to the next point the results of our estimation help us illustrate: the ability to measure dynamic changes in traits that are directly unobservable.

Observing changes in the levels of CSR over time

Of course, 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. Again, it is important to remember that the space our IRT Responsibility scores inhabit is one that already accounts for such dynamic changes.

To highlight the explicit incorporation of time in our model, we present the IRT Responsibility scores of selected major firms over time in Figure 2.

[Figure 2 about here.]

The solid black lines in Figure 2 illustrate our IRT Responsibility scores, while the grey shaded areas represent confidence bands for them. Dashed black lines in Figure 2 illustrate KLD Index values.

[18]Thinking about these firms in 2005 also helps assess the validity of our IRT Responsibility score. Despite many analyst's priors that Walmart was a relative laggard at CSR, which may have been true in 1991, the firm in 2005 demonstrated exceptional levels of social responsibility in ways that no other firm (or even the federal government) could in leveraging its supply chain to aid Hurricane Katrina victims (e.g., see Diermeier (2011) and Muller and Kräussl (2011)). On the other hand, thinking about Apple in 2005 paints quite a different picture, despite details on its activities not emerging until much later; that was one of the first few years that the firm began working with Foxconn in China, a firm at which labor and health conditions were dubious, among other ethical and social concerns (Duhigg and Barboza, 2012), and one of the first few years that the firm started engaging in aggressive maneuvers to avoid paying taxes in the United States (Duhigg and Kocieniewski, 2012)—both of which many analysts believe are highly questionable activities on social responsibility grounds (see, e.g., Christensen and Murphy (2004), Amaeshi et al. (2008), and Dowling (2013))

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 IRT Responsibility score and in KLD Index values are comparable—and we observe some meaningful differences on that front. The figure demonstrates that many notable firms exhibit marked improvements over time in our IRT Responsibility scores, while the same is not necessarily true for KLD Index values—bringing into question the validity of KLD Index values when considering the actual circumstances at many of these firms. With respect to our IRT Responsibility 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 IRT Responsibility score: the firm begins from a very low IRT Responsibility score—less than zero, 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 competitor, Target, begins with a much higher IRT Responsibility 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 score on our IRT Responsibility score exceeds Target's (with an 0.87 probability in the full simulated distribution). Importantly, had we looked at the KLD Index data 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 data for these firms.

The improvement trend for Walmart and Target in the IRT Responsibility scores is not necessarily the case for all retailers. We include two notable clothing retailers for bargain-minded shoppers: TJ Maxx and Vanity Fair. Both have very low IRT Responsibility scores early 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. For these two retailers, the trend in the KLD Index value is also relatively flat, although it is harder to say anything about the level of uncertainty in their trends given the lack of error bands.

While many notable 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. Notably, these downwards shifts in the IRT Responsibility score at Apple correspond to periods when founder & sometimes CEO Steve Jobs returns to the firm from temporary hiatuses–suggesting that theories about top management driving CSR might be supported in analyses using our IRT Responsibility score despite being challenged by measurement error issues when using the KLD Index (e.g., Hemingway and Maclagan, 2004; Hong and Minor, 2013).

Another trend to note from this view of the IRT Responsibility scores is that many firms, particularly those at the high end of the spectrum, demonstrate slight downturns toward the end of the data's time span (i.e., in the 2009-2012 period)—this is the case for industry leaders like IBM and GM. 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 et al., 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 IRT Responsibility score over time in Figure 3, Panel (a). We also plot median of the KLD Index over time in Figure 3, Panel (b).

[Figure 3 about here.]

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 panel (a) that the median firm's IRT Responsibility was on the rise. Were we to consider all firms' IRT Responsibility scores after 2001 in panel (a), we would infer that firms generally became less socially responsible. Upon 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 panel (b) over the same time period, we do not find the same trends. In the KLD Index, all firms, including those 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 like CSR and activist demands (Baron, 2001; Eesley and Lenox, 2006; Baron and Diermeier, 2007; Reid and Toffel, 2009). This finding in our IRT Responsibility score—but not in the KLD Index—might also speak to the claim that larger firms are generally less financially constrained than small firms and hence more free to spend on CSR initiatives (Hong et al., 2012).

Developing a more nuanced understanding of underlying CSR policies

The differences between the KLD Index and the IRT Responsibility score require further probing, given the several ways we have already seen them differ. In Figure 4, we create a scatterplot with the KLD Index on the horizontal axis and the IRT Responsibility score on the vertical axis.

[Figure 4 about here.]

Each dot in the figure represents a firm-year observation comparing how the IRT Responsibility scores measure up again the KLD Index values. For emphasis, earlier time points are depicted with darker dots. We also use a localized regression (loess) smoother to highlight overall trends in solid black–and include a diagonal dashed line that approximates the relationship we would expect if there was close to a one-to-one correspondence between the two measures. Generally, there is not such a correspondence. In fact, the overall correlation between the KLD Index and our measure is only .195.

Interestingly, only in the range from zero and up do the KLD Index values roughly track the IRT Responsibility scores. Our results suggest that many firms that one would label as irresponsible using only the KLD Index (because they have scores less than zero) are, in fact, average to above average in our IRT Responsibility score.[19]

Toward the middle of the KLD Index (i.e., near zero on the horizontal axis) we see many newer, smaller firms that our IRT Responsibility score labels as relatively irresponsible. Partially because our IRT Responsibility score is a continuous measure rather than an ordinal one, we see that there is an enormous amount of heterogeneity among firms with the same KLD Index value. Consider those observations with a KLD Index value of zero, of which there are 10,894. On the surface, it seems odd that over 25 percent of firms could be given the same score. It is even odder to think of these firms as being equivalent when recognizing that there are multiple ways to get to zero; 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 zero. In fact, the idea that firms may being doing things simultaneously that are socially responsible and socially irresponsible has confounded CSR researchers to date (e.g., Strike, Gao, and Bansal, 2006; Minor and Morgan, 2011; Kotchen and Moon, 2012).

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 zero? Our IRT Responsibility 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 IRT Responsibility score (approximately -8) among the KLD Index zeroes, whereas Ford has the highest IRT Responsibility score (approximately 10.5) among the same set of firms.

Hence, our theoretically motivated, explicitly dynamic approach allows the analyst to make such discriminations, which represents a large improvement over the baseline in which 25 percent of all firms are treated as if they were CSR neutral when, in fact, there is more heterogeneity in the data. The reason the IRT Responsibility measure is able to make a distinction between these firms is

[19]This finding is highly consistent with that in Delmas and Blass (2010) who look carefully at the rankings of 15 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. Hence, the more nuanced treatment of the underlying KLD data in the IRT modelling approach allows us to distinguish between multiple firms have the exact same KLD Index values, but that are in fact quite different.

To assess whether or not these differences between our IRT Responsibility 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 papers that make 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 critiques 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 authors claim that the KLD Index is too generous to some firms and too harsh to other firms, naming some firms explicitly or otherwise making claims about industries as a whole. Hence, we can compare these authors' claims about certain firms' treatment in the KLD Index to their treatment in the IRT Responsibility scores to assess the validity of our IRT-based measurement model.

[Figure 5 about here.]

Figure 5 shows scatterplots of the relative rankings of certain sets of firm-year observations on our IRT Responsibility score and on the KLD Index. 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 IRT Responsibility scores. Subplots group firms under headers about which Entine made specific predictions. For all six of the firm clusters depicted in Figure 5, Entine (2003) predicted that the KLD Index rated them too generously, although the reason why varied. In one instance, he predicted that financial firms and technology firms were rated too highly given the secretive nature of their businesses; the corresponding results show that the majority of our IRT Responsibility score predictions are consistent with this idea despite a few outliers. The Entine (2003) prediction holds slightly better

for technology firms (68% consistent) than financial firms (56% consistent).[20]

As another example, Entine (2003) predicted that Ben & Jerry's was given too generous of a score in the KLD Index for "their founders' mouthed anticorporate rhetoric" despite realities on the ground; in each firm-year observation for the firm, we find that our results are consistent with his. Entine (2003) also suggests that Microsoft may be treated too generously in the KLD Index given antitrust issues; the relative ranks that Microsoft receives are fairly similar under either than KLD Index or our IRT Responsibility score, suggesting that antitrust concerns either aren't a big issue or that Microsoft is ranked more-or-less appropriately. Entine (2003) also argues that firms with underfunded pensions will be rated too generously given this missing CSR issue in the KLD data collection effort.[21] Like with Microsoft and its antitrust issues, underfunded pensions do not seem like something that is a critical omitted policy, but even if it were, it could easily be incorporated into the analysis.

Finally, Entine (2003) suggests that the KLD Index, or a better measure of CSR, should be able to predict public scandals at firms including those that occurred at Anderson Accounting, Enron, WorldCom, Adelphia, Tyco, and Tenet Healthcare; this is the one area where our IRT score and Entine (2003)'s predictions do not accord well at all. Stepping back, however, it is a lot to ask of a measure of a latent trait to pick up something like this, when management's actions may be intentionally oriented towards not getting caught, so that these firms might have choose to mimic "good firms" while doing "bad." Overall, the analysis in Figure 5 suggests that our IRT Responsibility measure of CSR has solid validity.

Entine (2003) and Delmas and Blass (2010) come together in suggesting that it is particularly hard for the KLD Index to rate firms in industries where the opportunities to avoid environmental degradation are rare, but where the positives are difficult to observe. Entine (2003) argues that all of these firms are rated too harshly in the KLD Index. Delmas and Blass (2010) argue that some firms are

[20] In the figure, we include the following firms that are S&P 500 members in the appropriate sector: financials (AFLAC, Allstate, American Express, AIG, Bank of America, BB&T, Blackrock, Citigroup, Fifth Third Bancorp, Genworth Financial, JP Morgan, KeyCorp, Legg Mason, Lincoln National, MetLife, Moody's, PNC Financial, Stat Street, T Rowe Price, Wells Fargo) and technology (Adobe, Advance Micro Devices, Apple, Applied Materials, Broadcom, Cisco Systems, Electronic Arts, Intel, Intuit, JDS Uniphase, KLA-Tencor, Motorola, SanDisk, Segate, Symantec, Teradyne, Western Digital).

[21] Specifically, Entine (2003) names the following firms for having underfunded pension issues: Verizon, Cummins Engine, American Airlines, and Delphi Automotive.

treated too harshly (along the lines of Entine (2003)) while others are treated too generously. Figure 6 illustrates the analysis for this set of firms.

[Figure 6 about here.]

As in Figure 5, Figure 6 shows the relative rankings of firms on our IRT Responsibility score and on the KLD Index. The dots, their coloration, and the 45-degree line all maintain the same interpretation. First, looking at the more generic critiques Entine (2003) makes for the energy, resources, and chemicals industries, we see that the IRT Responsibility scores overwhelming produce results consistent with his expectations.[22] That is, we agree with his assessment that these firms may have been rated too harshly given structural issues that make them polluters with problems that are difficult to solve. Moving on to Delmas and Blass (2010)'s predictions—in which some firms are rated too harshly and others too generously—we find evidence that is even more devastating to the validity of the KLD Index measure, largely favoring the validity of our IRT Responsibility score measure.

The direct comparisons between the KLD Index and the IRT Responsibility score highlights two final important points about IRT models:

3. Relative to the additive KLD Index, a Bayesian IRT analysis offers a much more nuanced (and different) picture of firms, especially for firms that have a large number potentially "offsetting" strengths and concerns.
4. The IRT Responsibility scores reflect a number of realities that critics of the KLD Index have noted in specific examples; this is because the Bayesian technique does not treat every underlying CSR indicator equally but instead uses the overall structure of the data to determine their relative importance.

Predictive capabilities of IRT Responsibility scores vs. KLD Index

Our discussion thus far demonstrates in several ways how IRT Responsibility scores are superior to the KLD Index on theoretical and substantive grounds, but some readers may also be interested in these scores' relative predictive success. Perhaps the toughest test of predictive success is whether

[22] The firms included in the three graphs are: for energy (Baker Hughes, Chevron, Conoco Philips, CONSOL Energy, Diamond Offshore Drilling, EOG Resources, ExxonMobil, Halliburton, Hess, Marathon, Nobel Newfield Exploration, Phillips 66 Tesoro, Valero, and Williams Cos.), for natural resources (US Steel, 3M, Corning, International Paper, Nucor, Newmont Mining, Freeport McMoran), and for chemicals (Dow, DuPont, Eastman Chemical, Ecolab, FMC Corp, PPG Industries, Sigma-Aldrich).

IRT Responsibility scores do a better job than the KLD Index itself of predicting behavior on new indicators for CSR "strengths" or CSR "concerns" that become components of the KLD Index.

To conduct two head-to-head prediction battles, we utilized some interesting opportunities that arise from the KLD data. The first involves an addition to the dataset in 1999 (a new variable for concerns about a firm's handling of climate change controversies such as dependence on coal) and the second involves an addition to the dataset in 2010 (a new variable for concerns about governance structures). To leverage the first of the additions, we ran the Bayesian IRT routine on the KLD data through 1998, thus developing a new set of scores that did not incorporate information about the *ex-post* addition of the new indicator. We then ran two bivariate probit regression models with the 1999 climate change metric as the dependent variable: one with the 1998 IRT Responsibility score as the lone explanatory variable, and another with the 1998 KLD index as the lone explanatory variable. We repeated the same process to predict the new 2010 variable for undesirable governance structures using the 2009 KLD index and a 2009 IRT Responsibility score based on the data through 2009.

To assess how well the two responsibility indices predict the new KLD variables, we assess 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 consider how these results turn out across *all* possible selections of c. This is just the sort of analysis we conducted.

We summarize these results using a powerful graphical tool known as a receiver operating characteristic (ROC) curve in Figure 7. 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.

[Figure 7 about here.]

For the 1999 observation of climate change issues in the left-hand panel of Figure 7, the area under the IRT Responsibility score curve (0.684) is greater than the area under the KLD Index curve (0.650), indicating that the former has greater predictive capabilities than the latter. We can more readily see that the IRT Responsibility score predicts better than the KLD Index for swaths of false positive rates, most notably at very low levels. This range is especially important because many of the predicted probabilities for the climate change variable are quite low.

For the 2010 observation of governance issues in the right-hand panel of Figure 7, the superior predictive performance of the IRT Responsibility score is much more obvious throughout; this time, the IRT responsibility score has an area under the curve of 0.767, compared with the KLD Index's area under the curve of 0.524. Notably, at high false positive rates, the KLD Index performs *worse* than random guessing does. As a robustness check to ensure that the IRT Responsibility score is not simply taking advantage of the 2010 data's larger and more heterogeneous set of firms, we constructed the same ROC curve for the 2010 analysis using only S&P and Domini firms. The story remains the same: the IRT Responsibility score dominates the KLD Index for predictive power, and the latter is again sometimes worse than random guessing.

How can we explain the marked advantage of IRT over KLD in 2010 to the less pronounced (but still noticeable) advantage in 1999? The answer lies in the different methodological approaches underlying the IRT score and the KLD Index. Arguably, the KLD index should perform relatively better in 1999 because the dataset at that point in time only includes large, relatively homogenous firms from the S&P and Domini indices, where the mapping from firm behavior to CSR is relatively

more straightforward. The IRT model works well then, but works even better later, i.e., in 2010, as it "learns" from the addition of data in a way that an additive index does not. As new CSR items and new firms are added to the dataset, the model is able to perform better since it has more information to work with. Meanwhile, the additive index suffers from the inclusion of new traits, since the equal weight assumption becomes even less tenuous as indicators change over time. On top of this, the continuity of the measure allows the simulation algorithm to place firms along the scale with an amount of nuance that allows for meaningful differentiation within and across clusters of similar firms, which is especially helpful as the firms become more heterogeneous after 2001 when more firms outside of the S&P 500 are added to the dataset.

CONCLUSION

In this paper, we have demonstrated the usefulness of Item Response Theory modeling for strategic management researchers by applying it to commonly used corporate social responsibility data. 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. Furthermore, they provide a better sense of how reliable the measures are. Our analysis shows that the existing additive indices using KLD data sometimes overstate a firm's CSR levels, and sometimes understate it, often in unexpected ways. Our analysis also shows that some firms are easier to distinguish on CSR 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 IRT Responsibility scores produce a more nuanced measure of CSR than the KLD Index. This is most vividly demonstrated by looking at the big differences in IRT Responsibility scores for firms that receive identical KLD Index scores of 0.

We conclude the paper by discussing the implications of our approach for CSR research as well as strategic management research more broadly. Our paper contributes to the CSR literature in three ways. First, the data we generated in this paper opens up new avenues of inquiry. We have not only shown that the traditional measure of CSR, based on an additive index, is problematic, but also offered a solution. That solution is an IRT-based measure, the IRT Responsibility score, which lays the foundation for new empirical work in CSR that takes advantage of this new score, as well as work

revisiting published results that utilized an additive KLD measure. In the paper, we show how our basic model can assess previous critiques about the KLD measure—that it is too generous to some firms and too harsh with others—and speak to ongoing debates in the literature.

Second, because the IRT framework can incorporate additional information into the statistical estimation of CSR beyond items in the KLD Index, the application-specific measures of IRT modeling in CSR become readily implementable. For instance, researchers interested in learning whether environmental regulations influence levels of CSR could incorporate these rules into the model. An analyst may also want to determine whether there is more than one "dimension" to CSR. 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 which potential inflict social harm represent a different dimension than actions which 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 CSR as a construct is empirically valid.

Third, in this paper we have focused on the 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? The differences in the scores between our measure and the standard additive index implies that all KLD items are not created equal.

Our paper is useful, then, for researchers who want to use our data for their own work or to revisit existing work, adapt the IRT model for new applications, or explore the underlying items comprising CSR. Of course, not all readers of *SMJ* intend to work in the area of CSR, but this paper's reach extends far beyond this one research area. As we noted at the outset of this paper, several key measures in management, including those for corporate governance and entrepreneurial orientation, are constructed based on a set of items or actions, and therefore could be improved by the application of IRT.

Boyd, Gove, and Hitt (2005b, 367), in tackling measurement error problems in the debate over whether diversification by firms is due to agency costs, write, "Our results provide strong evidence that

the debate between authors is largely an artifact of measurement error." IRT models have the potential to act as an arbiter of competing claims across many literatures. In the area of corporate governance, for instance, the IRT model could address which parts of the "G-index" should be part of a measure of corporate governance, and which should not. Researchers with theories about which types of firms, for instance, ought to have a board chair separate from the CEO and which firms are best governed with unified chairs, could build those beliefs into their IRT measurement model. Various other governance debates could be revisited with an IRT-based "G-index."

Our paper also has implications for researchers seeking to create *new* measures of a phenomenon (perhaps even CSR) from scratch. Though there will always be disagreement about which items should and should not be part of the construction of a given measure, the IRT model can help sort out competing claims rather than relying on intuition or guesswork (though both could be important inputs into the IRT model via the "priors"). The result will be more reliable indicators upon which important empirical analyses of key phenomenon can be built.

Of course, IRT modeling, like all methods, has its limitations. It is computationally demanding and requires some programming knowledge, though recent software advances make the nature of this programming more manageable than it had been previously. Also, the quality of the output is only as good as the quality of the data input and associated theoretical model. The statistical model's heft should not lead to complacency in other aspects of measurement.

The overall message of this paper, however, 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—strikes us as a worthy investment.

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.

Albert JH. 1992. Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics*, **17**(3):251–269.

Albert JH, Chib S. 1993. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, **88**(422):669–679.

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.

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.

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

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

Bebchuk LA, Cohen A, Ferrell A. 2009. What matters in corporate governance? *Review of Financial Studies*, **22**(2):783–827.

Bhagat S, Bolton BJ, Romano R. 2008. The promise and peril of corporate governance indices. *Columbia Law Review*, **108**(4):1803–1882.

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

Bock RD, Gibbons R, Muraki E. 1988. Full-information item factor analysis. *Applied Psychological Measurement*, **12**:261–280.

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

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

Boyd BK, Gove S, Hitt MA. 2005b. Consequences of measurement problems in strategic management research: the case of Amihud and Lev. *Strategic Management Journal*, **26**(4):367–375.

Burris V. 2001. The two faces of capital: corporations and individual capitalists as political actors. *American Sociological Review*, **66**(3):361–381.

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 & Society*, **38**(3):268–295.

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

Chin M, Hambrick DC, Trevino LK. 2013. Political ideologies of CEOs: the influence of executives' values on corporate social responsibility. *Administrative Science Quarterly*, **58**:197–232.

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. 1989. Strategic management of small firms in hostile and benign environments. *Strategic Management Journal*, **10**(1):75–87.

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. 2003. Corporate governance: decades of dialogue and data. *Academy of Management Review*, **28**(3):371–382.

Das J, Hammer JS. 2004. Which doctor? combining vignettes and item response to meaure doctor quality. World Bank Paper, accessed 7/25/13 at elibrary.worldbank.org/content/workingpaper/10.1596/1813-9450-3301.

deBakker 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 & Society*, **44** (3):283–317.

Delmas M, Blass VD. 2010. Measuring corporate environmental performance: the trade-offs of sustainability ratings. *Business Strategy and the Environment*, **10**(4):245–260.

Delmas M, Etzion D, Nairn Birch N. Forthcoming. Triangulating environmental performance: what do corporate social responsibility ratings really capture? *Academy of Management Perspectives*.

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

DiMaggio PJ, Powell WW. 1983. The iron cage revisited: institutional isomorphism and collective rationality in organizational fields. *American Sociological Review*, **48**(1):147–160.

Dowling GR. 2013. The curious case of corporate tax avoidance: is it socially irresponsible? *Journal of Business Ethics*, **Advance Online Publication**.

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

Duhigg C, Kocieniewski D. How Apple sidesteps billions in taxes. *New York Times*, 2012. April 28.

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 the 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.

Frederick WC. 1994. From CSR1 to CSR2. *Business & Society*, **33**(2):150–164.

Fremeth A, Richter BK, Schaufele B. 2013. Campaign contributions over CEOs' careers. *American Economic Journal: Applied Economics*, **5**(3):170–188.

Gelman A, Carlin JB, Stern HS, Rubin DB. 2003. *Bayesian Data Analysis*. Chapman & Hall, Boca Raton, FL, second edition.

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

Godfrey PC, Hill CW. 1995. The problem of unobservables in strategic management research. *Strategic Management Journal*, **116**: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.

Gompers PA, Ishii JL, Metrick AM. 2003. Corporate governance and equity prices. *Quarterly Journal of Economics*, **118**(1):107–156.

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

Griffin JJ, Mahon JF. 1997. The corporate social performance and corporate financial performance debate: twenty-five years of incomparable research. *Business & 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. ; 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 B, Minor D. 2013. Good (bad) company or good (bad) manager? exploring the antecedents of CSR. Northwestern University Working Paper.

Hong H, Kubik JD, Scheinkman JA. 2012. Financial constraints on corporate goodness. *NBER Working Paper 18476*.

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.

Jackman S. 2001. Multidimensional analysis of roll call data via Bayesian simulation: identification, estimation, inference, and model checking. *Political Analysis*, **9**(3):227–241.

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

Kacperczyk AJ. 2009. With greater power comes greater responsibility: takeover protections and corporate attention to stakeholders. *Strategic Management Journal*, **30**(3):261–285.

Kamata A, Bauer DJ. 2008. A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling*, **15**:136–153.

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.

KLD Stats. 2008. Getting started with KLD STATS and ratings definitions. Working Guide.

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

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

Lord FM, Novick MR. 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.

Lumpkin GT, Dess GG. 2001. Linking two dimensions of entrepreneurial orientation to firm performance: the moderating role of environment and industry life cycle. *Journal of Business Venturing*, **16**(5):428–451.

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 & Management Strategy*, **20**(1):3–41.

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

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 on SSRN at `http://ssrn.com/absract=1866371`.

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 & Society*, **45**(1):20–46.

McWilliams A, Siegel DS, Wright PM. 2006. Corporate social responsibility: Strategic implications. *Journal of Management Studies*, **43**(1):1–18.

Minor D. 2013. The value of corporate citizenship: protection. Northwestern University Working Paper.

Minor D, Morgan J. 2011. CSR as reputation insurance *Primum Non Nocere*. *California Management Review*, **53**(3).

MSCI. 2012. MSCI ESG stats: user guide & esg ratings definition. *ESG Research*, 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. 2011a. Strategic corporate social responsibility and environmental sustainability. *Business & Society*, **50**(1):6–27.

Orlitzky M, Siegel DS, Waldman DA. 2011b. Strategic corporate social responsibility and environmental sustainability. *Business & Society*, **50**(1):6–27.

Poole KT, Rosenthal H. 1991. Patterns of congressional voting. *American Journal of Political Science*, **35**(1):228–278.

Quinn KM. 2004. Bayesian factor analysis for mixed ordinal continuous responses. *Political Analysis*, **12**(4):338–353.

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.

Richard OC, Barnett T, Dwyer S, Chadwick K. 2004. Cultural diversity in management, firm performance, and the moderating role of entrepreneurial orientation dimensions. *Academy of Management Journal*, **47**(2):255–266.

Rowley T, Berman S. 2000. A brand new brand of corporate social performance. *Business & 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.

Sonnenfeld J. 2004. Good governance and the misleading myths of bad metrics. *Academy of Management Executive*, **18**(1):108–113.

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.

Takane Y, De Leeuw J. 1987. On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, **52**(3):393–408.

Thurstone LL. 1925. A method of scaling psychological and educational tests. *Journal of Educational Psychology*, **16**:433–451.

Treier S, Jackman S. 2008. Democracy as a latent variable. *American Journal of Political Science*, **52**(1):201–217.

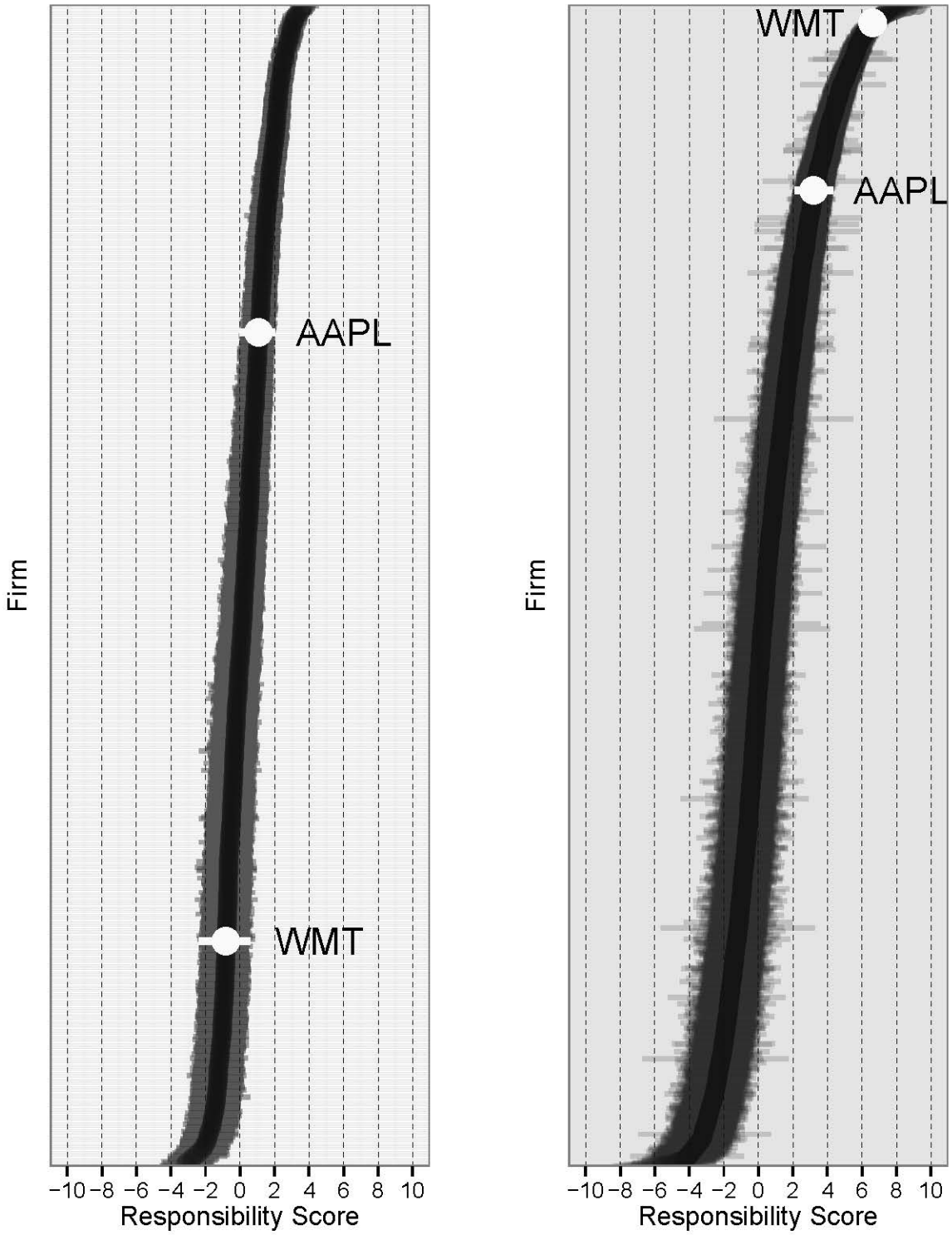
Vanden Bergh RG, Holburn GL. 2007. Targeting corporate political strategy: theory and evidence from the US accounting industry. *Business and Politics*, **9**(2):1–31.

Venkatraman N, Grant JH. 1986. Construct measurement in organizational strategy research: a critique and proposal. *Academy of Management Review*, **11**(1):71–87.

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*, **15**(3):287–296.

Walls JL, Phan PH, Berrone P. 2011. Measuring environmental strategy: construct development, reliability and validity. *Business & Society*, **50**(1):71–115.



(a) 1991

(b) 2005

Figure 1: All firms in two years.

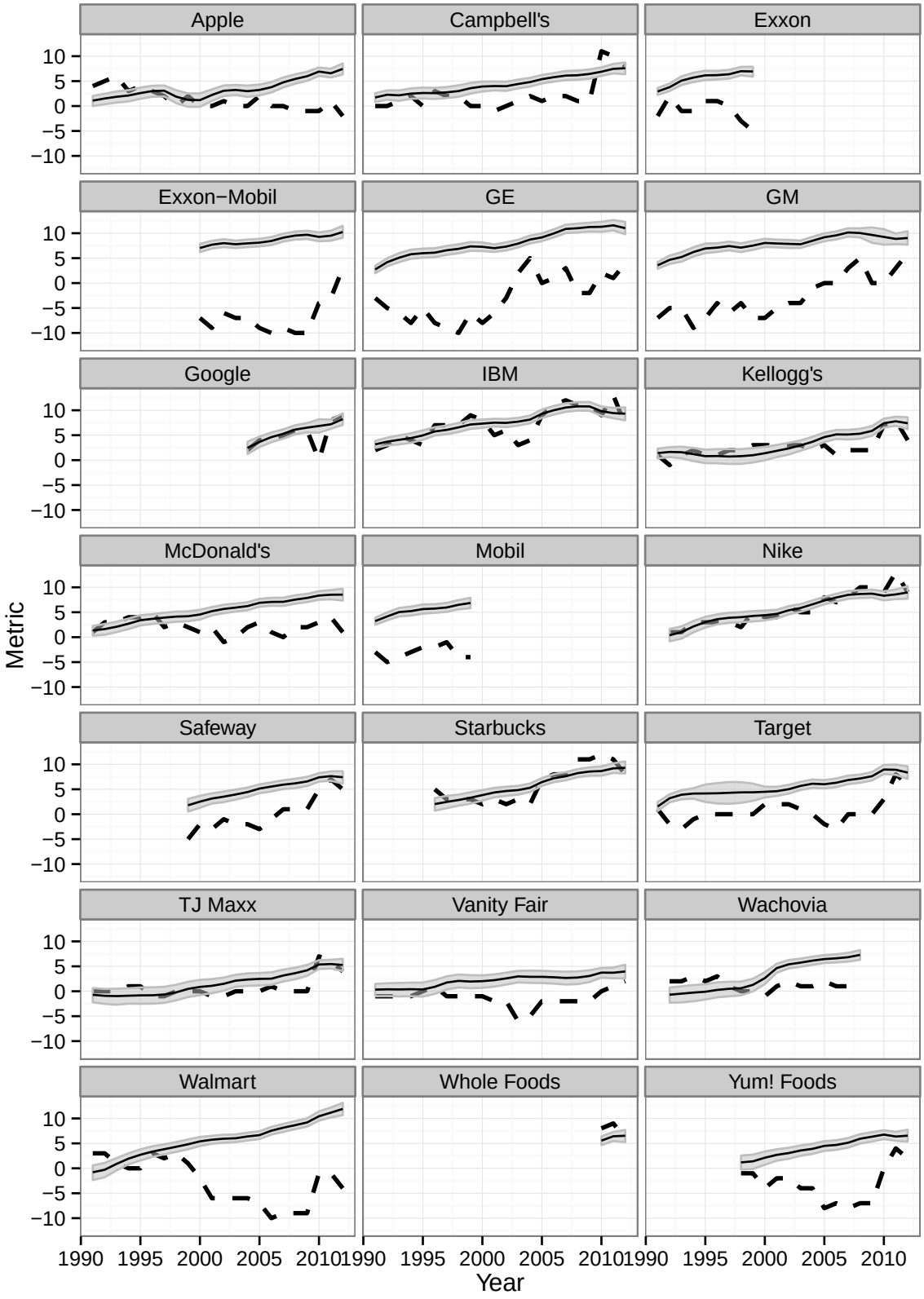
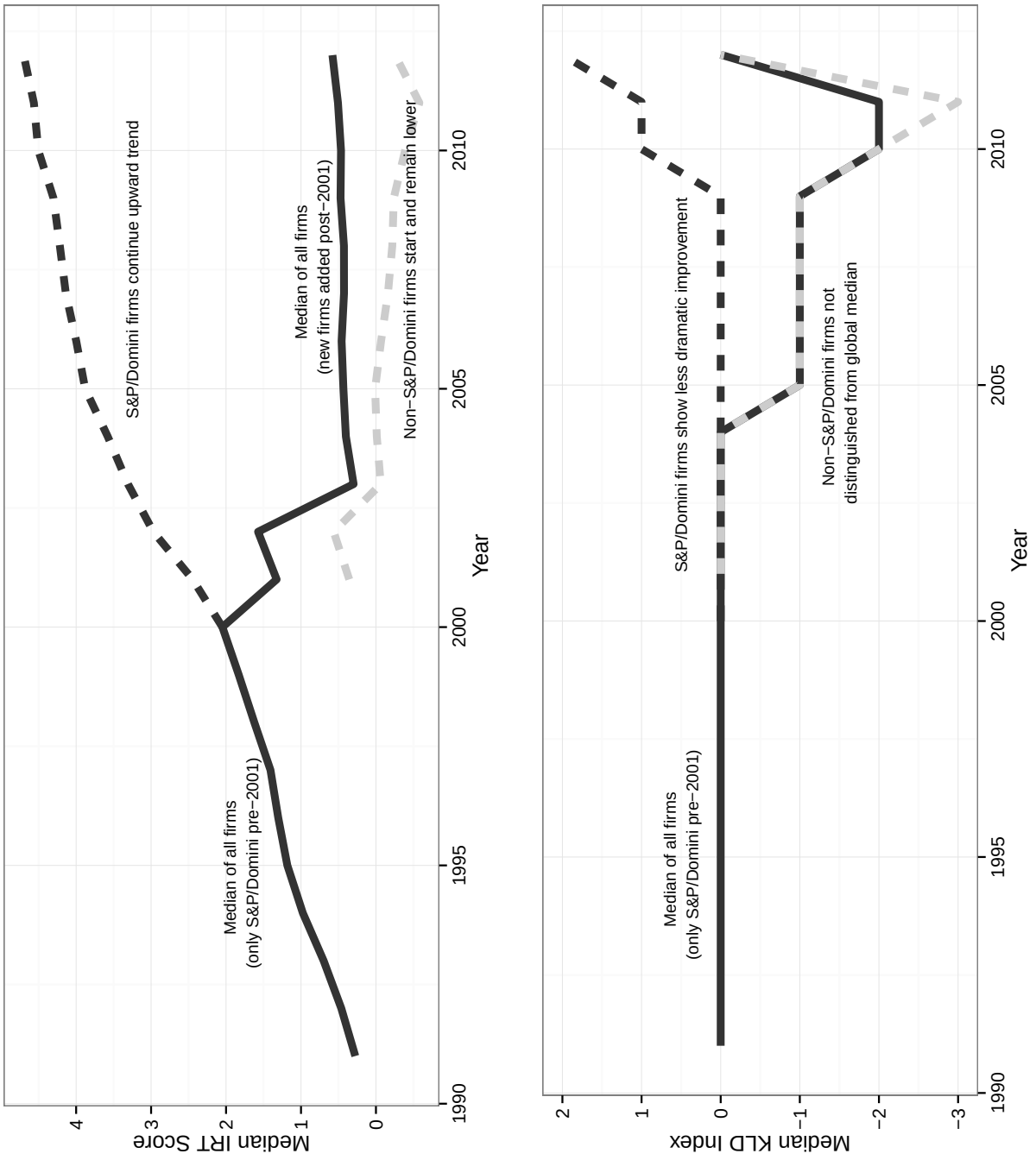


Figure 2: Select firms over time. Bayesian IRT Responsibility score shown with solid line with confidence interval. KLD Index shown with dashed line.



(a) IRT Medians

(b) KLD Index Medians

Figure 3: Median scores over time.

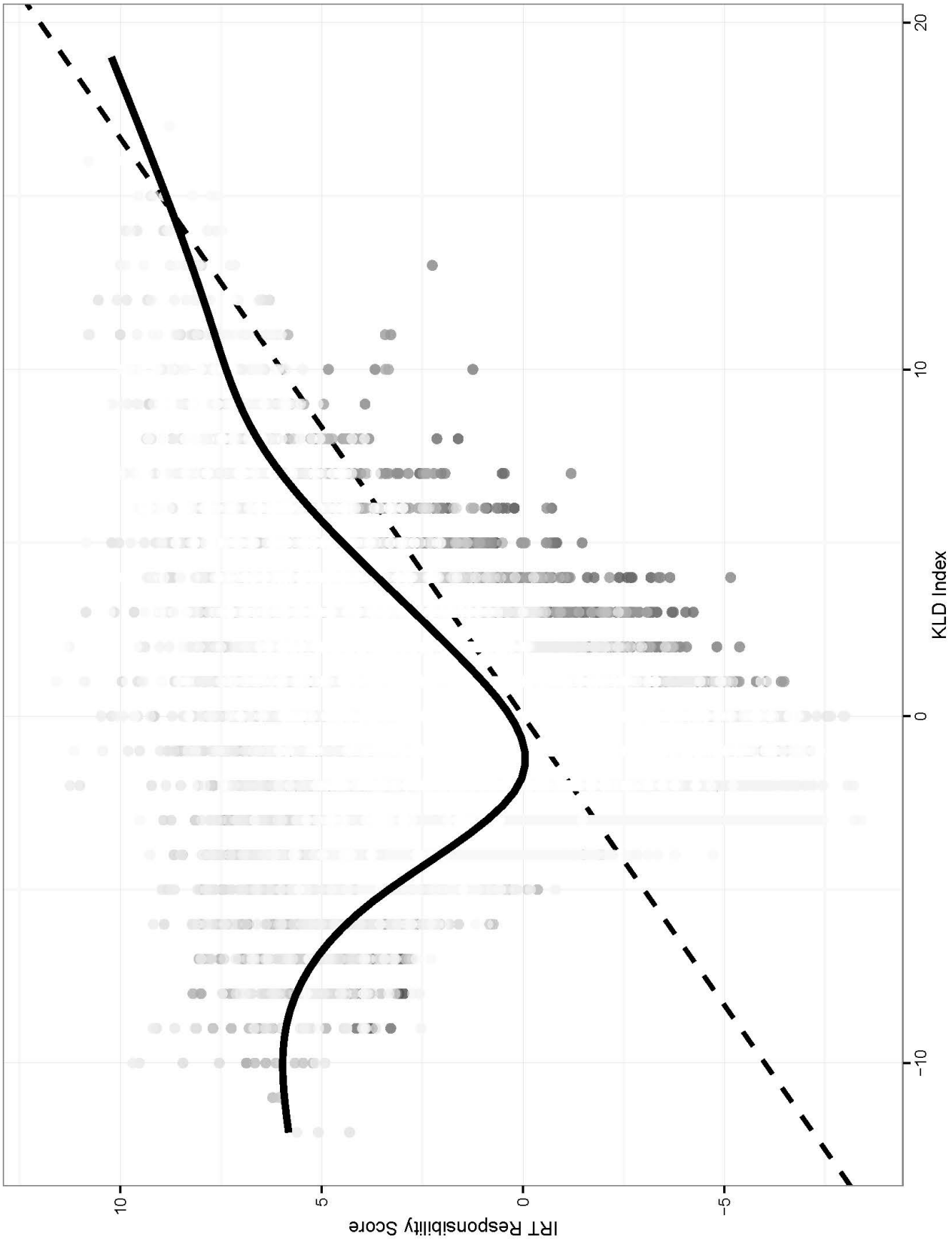


Figure 4: KLD Index versus IRT Responsibility score by time (earlier timepoints are darker).

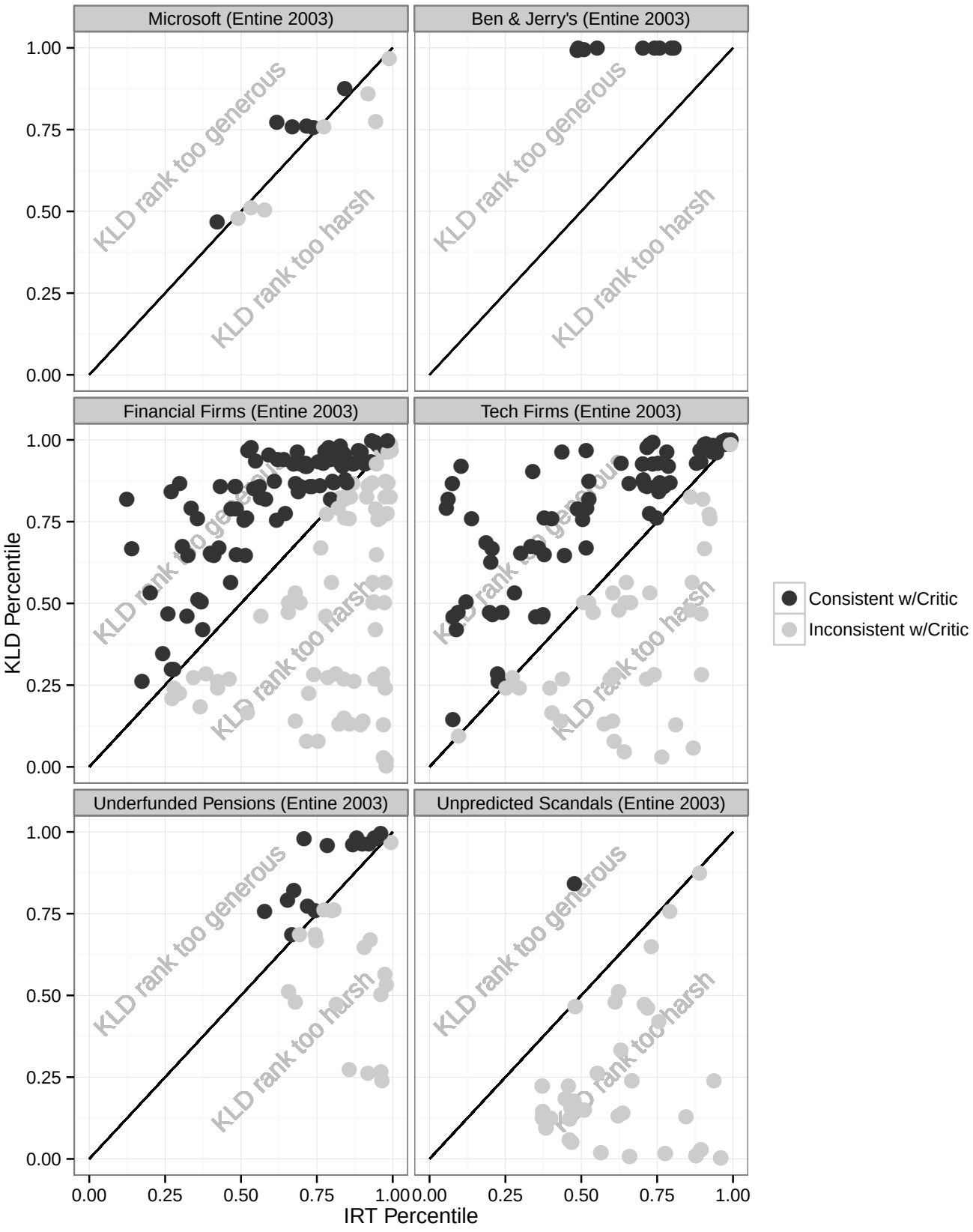


Figure 5: Relative rankings, IRT Responsibility scores versus KLD Index, Entine (2003) firms.

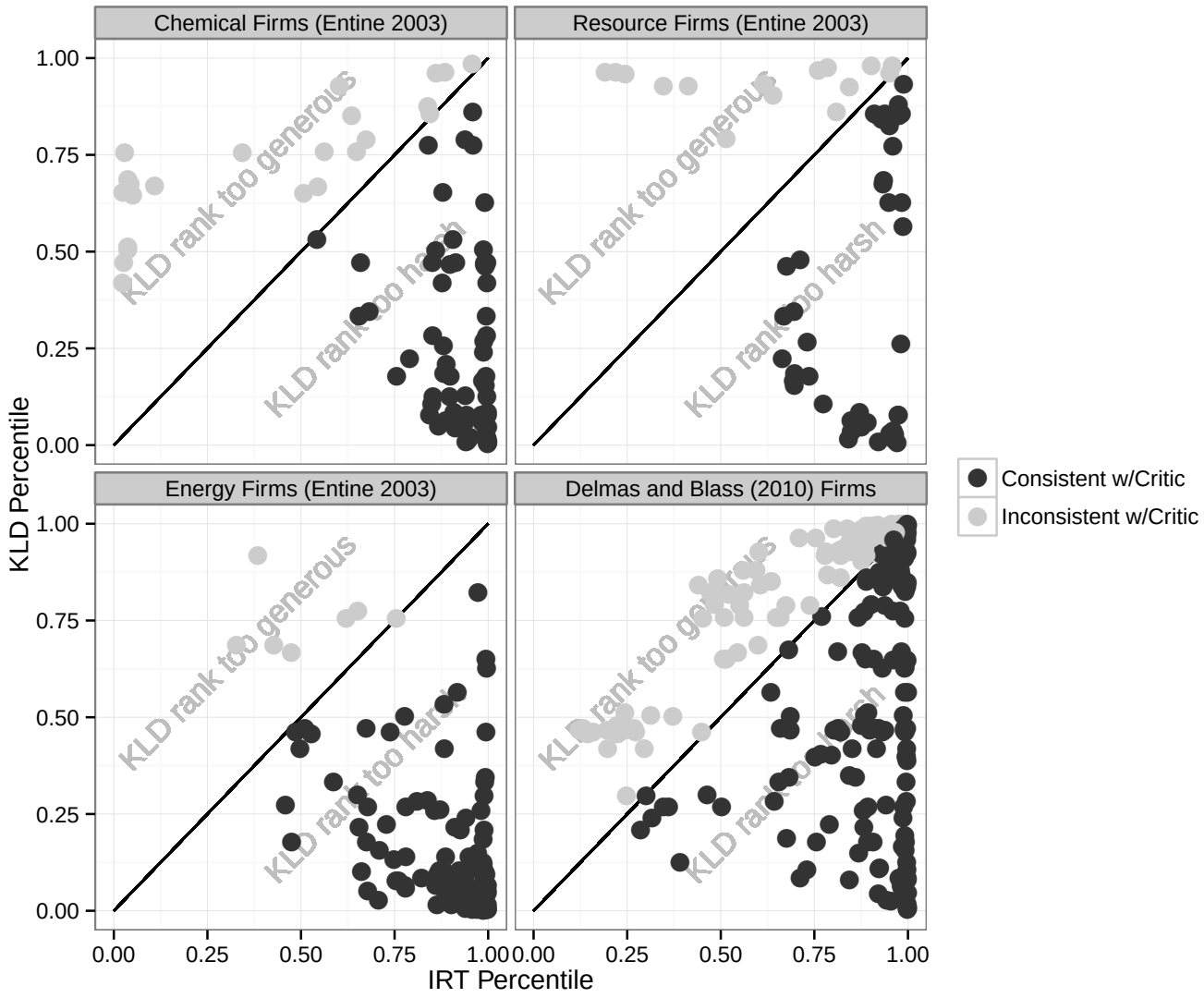


Figure 6: Relative rankings, IRT Responsibility scores versus KLD Index, environmental-impact firms.

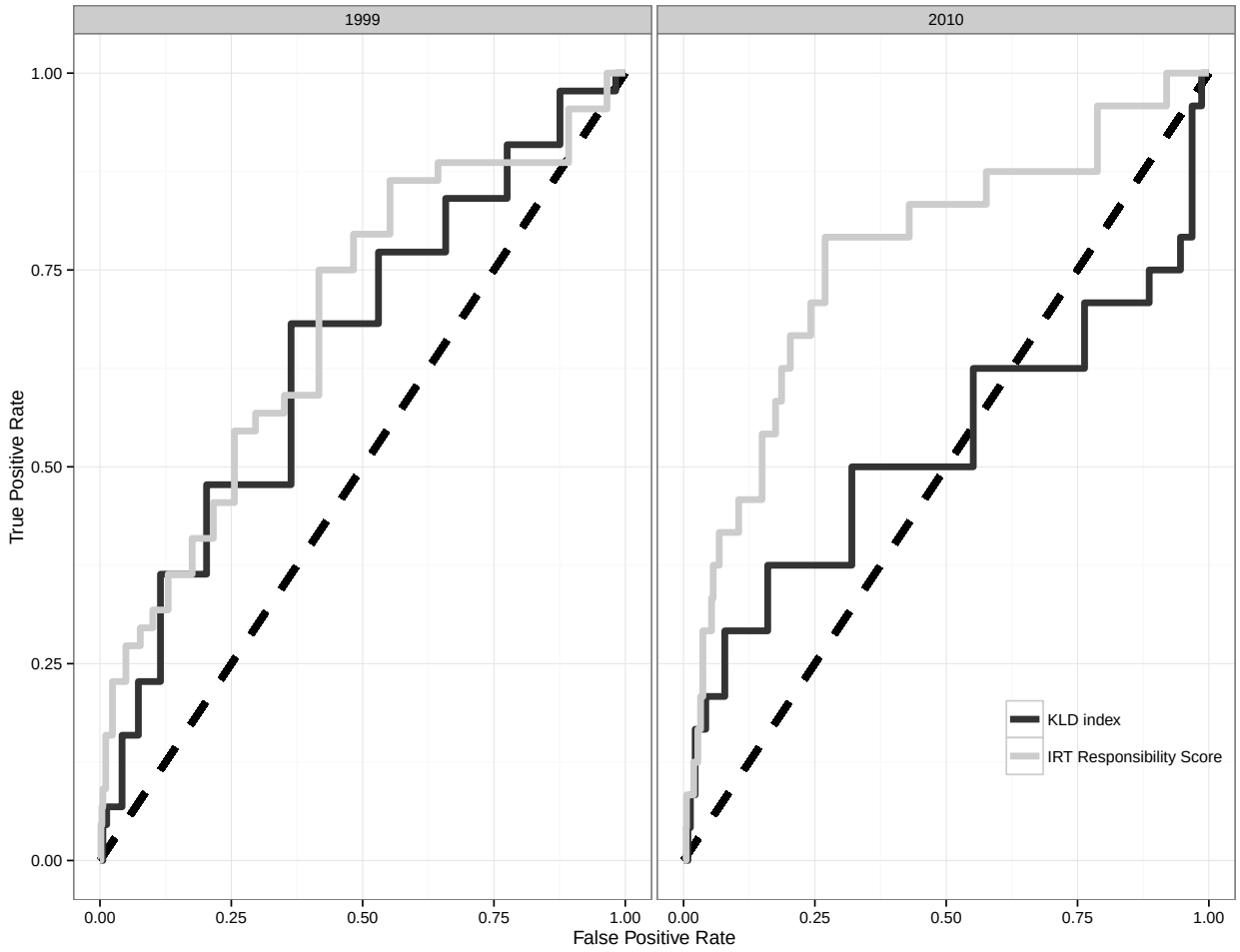


Figure 7: ROC Curves summarizing Relative Predictive Power of the KLD Index versus IRT Responsibility scores.

Appendix for
Using Item Response Theory to Improve Measurement in Strategic Management Research: An Application to Corporate Social Responsibility

March 25, 2014

In this appendix, we give a more complete account of our statistical model's assumptions and of the MCMC routine for posterior simulation. Our approach is similar to that of Martin and Quinn (2002). Indeed, their model is a special case of ours; with certain assumptions, the two approaches are identical. For the purposes of introducing the model to a new audience, and and to demonstrate how existing software can be used for estimation, we maintain as much similarity to Martin and Quinn (2002) as possible and highlight points of potential departure.

In the main body of the paper, the latent utility received by firm i for choosing to adopt policy j in time period t is represented by

$$z_{i,j,t} = \alpha_{j,t} + \beta_{j,t}\rho_{i,t} + \varepsilon_{i,j,t}$$

All of the terms except for $\varepsilon_{i,j,t}$ are discussed in the main body of the paper. Following common practice, we assume that $\varepsilon_{i,j,t}$ is identically and independently distributed with mean zero and variance one. Our aim is to estimate each α, β, and ρ terms. To discuss how we do so, we simplify notation. Collect the respective $\alpha_{j,t}$ terms for all the observables in all the time periods into a vector $\boldsymbol{\alpha}$, and do the same for the respective $\beta_{j,t}$ terms into a vector $\boldsymbol{\beta}$. Collect all the latent trait scores of all firms in all time periods into vector $\boldsymbol{\rho}$. Finally, collect all the decisions on all observables by all observed firms in all time periods into a vector $\boldsymbol{d}$. We adopt a Bayesian inference protocol, so our object of interest is

the posterior density

$$p\left(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\rho} \mid \boldsymbol{d}\right) \ \propto \ p\left(\boldsymbol{d} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\rho}\right) p\left(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\rho}\right).$$

Here the symbol $\propto$ denotes proportionality: the term on the right-hand side of the expression is the numerator familiar from Bayes' rule, where $p\left(\boldsymbol{d} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\rho}\right)$ is the *likelihood function* and $p\left(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\rho}\right)$ is the *prior distribution*. Thus the expression above states that the posterior distribution is proportional to the product of the likelihood of the decisions conditional on the underlying parameters and the prior belief about those parameters. For the likelihood function, the assumptions made about the error terms imply that the decisions have Bernoulli likelihood

$$p\left(\boldsymbol{d} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\rho}\right) \ \propto \prod_{t=1}^{T} \prod_{j \in J_t} \prod_{i \in I_j} \Phi\left(\alpha_{j,t} + \beta_{j,t}\rho_{i,t}\right)^{\delta_{i,j,t}} \left(1 - \Phi\left(\alpha_{j,t} + \beta_{j,t}\rho_{i,t}\right)\right)^{1-\delta_{i,j,t}},$$

where $\delta_{i,j,t} = 1$ if $D = A$, $\delta_{i,j,t} = 0$ if $D = R$, and Φ is the standard normal cumulative distribution function. Here J_t is the set of all observables in time period t and I_j is the set of all firms observed for observable j.

For identification purposes, the assumed prior distributions must be semiinformative. The literature has adopted a standard set of semiinformative prior for such purposes, which we adopt here. First, we are concerned with the change in IRT Responsibility scores over time and as such must model a mapping from one time period's IRT Responsibility score into the next. Firms' initial IRT Responsibility in (unobserved) time period 0, $\rho_{i,0}$, is distributed

$$\rho_{i,0} \sim \mathcal{N}\left(m_{i,0}, C_{i,0}\right),$$

where we specify $m_{i,0}$ and $C_{i,0}$ numerically below. With the anchor in place, a *random walk* process dictates how one period's responsibility score maps into the next period's:

$$\rho_{i,t} \sim \mathcal{N}\left(\rho_{i,t-1}, \Delta_{\rho_{i,t}}\right)$$

$\Delta_{\rho_{i,t}}$, the variance of the normal distribution from which a IRT Responsibility score is drawn, dictates how closely information from the previous period relates to information in the current period. If it is very small, then the time series of IRT Responsibility for a firm over time approaches a constant value. If it is very large, then the time series is essentially unrelated to itself across time. Martin and Quinn (2002) observe that this is a "happy median" between one extreme (not modeling changes over time at all) and the other (not allowing one time period's responsibility to be related to the next).

We assume that the difficulty and discrimination terms $\alpha_{j,t}$ and $\beta_{j,t}$ are drawn from a multivariate normal distribution,

$$\begin{bmatrix} \alpha_{j,t} \\ \beta_{j,t} \end{bmatrix} \sim \mathcal{N}_2 \left(\boldsymbol{b}_0, \boldsymbol{B}_0 \right),$$

for all observables across all time periods. Importantly, for these observable-specific terms, we do *not* model dynamic effects in policy specific attributes. Instead, we treat each observable "as a new case" in each year. To be sure, the analyst could utilize a random walk model like we utilize for the IRT Responsibility score. But, the marginal costs and benefits of adopting a given policy, which presumably fluctuate over time, are just the sort of thing we aim to learn more about. Put differently, we know that firms have sticky policies over time, but we *don't* know what CSR is. That is the very aim of the enterprise, and so allowing fluctuations seems the best approach for the learning process. Put differently, and in the terms stated above, we assume that the variance of a random-walk process underlying changes in difficulty and discrimination over time is *infinite*. This is an assumption that we make to maximize similarity to the extant IRT literature, but it can be relaxed—and time dependency in policy-specific parameters can be modeled—in future work.

The joint posterior distribution is simulated via Markov chain monte carlo (MCMC) methods. The algorithm, developed by Albert and Chib (1993), proceeds in three parts:

1. First, for all time periods, all firms, and all observables, we simulate $z_{i,j,t}$. Given the assumptions

of the model, the distribution of $z_{i,j,t}$ is given by

$$p\left(z_{i,j,t} \mid d_{i,j,t}, \rho_{i,t}, \alpha_{j,t}, \beta_{j,t}\right) = \begin{cases} \mathcal{N}_{[0,\infty]}\left(\alpha_{j,t} + \beta_{j,t}\rho_{i,t}, 1\right), & d_{i,j,t} = 1 \\ \mathcal{N}_{[-\infty,0]}\left(\alpha_{j,t} + \beta_{j,t}\rho_{i,t}, 1\right), & d_{i,j,t} = 0 \\ \mathcal{N}\left(\alpha_{j,t} + \beta_{j,t}\rho_{i,t}, 1\right), & d_{i,j,t} = NA. \end{cases}$$

That is, if $d_{i,j,t}$ is 1, the latent utility is distributed truncated normal at zero on the left; alternatively, if $d_{i,j,t}$ is 0, the latent utility is distributed truncated normal at zero on the right. For missing observations (which constitute a large proportion of our data), we make no truncation assumptions. Note that this step discriminates between observed and unobserved data: observed data are drawn from the appropriate truncated normal distributions, whereas unobserved data are drawn from the full normal distribution. Collect all of the latent utilities into a vector $\boldsymbol{z}$.

2. For all time periods and all firms, we simulate the $\alpha_{j,t}$ and $\beta_{j,t}$ terms. Let $\boldsymbol{\rho}^*_{i,t} = \begin{bmatrix} 1 \\ \rho_{i,t} \end{bmatrix}$, and let $\boldsymbol{\rho}^*_t$ be the $|I_j| \times 2$ matrix formed by stacking all these vectors for all firms. Then the observable-level parameters are distributed bivariate normal:

$$p\left(\begin{bmatrix} \alpha_{j,t} \\ \beta_{j,t} \end{bmatrix} \mid \boldsymbol{d}, \boldsymbol{z}, \boldsymbol{\rho}\right) = \mathcal{N}_2\left(\mathbb{E}\left[\left(\boldsymbol{\rho}^*_t\right)' \boldsymbol{z}_{i,t} + \boldsymbol{B}_0^{-1}\boldsymbol{b}_0\right], \left(\boldsymbol{\rho}^*_t\right)' \boldsymbol{\rho}^*_t + \boldsymbol{B}_0^{-1}\right).$$

3. For all firms in all time periods, simulate the CSR score. This is done via a forward-filtering, backward sampling algorithm described at length by Martin and Quinn (2002). They place the algorithm in the context of a general multivariate dynamic linear model—that is, one where each firm's ideal points constitute a time series—that allows for more complicated specifications of the general process of interest. Begin by rewriting the latent utility equation into a form we will call the *observation equation*:

$$z_{i,\cdot,t} - \boldsymbol{\alpha}_t = \boldsymbol{\beta}_t \boldsymbol{\rho}_{i,t} + \boldsymbol{\varepsilon}_{i,\cdot,t},$$

where $\cdot$ makes it explicit that we are considering a time period across all observables. For the dynamic ideal point estimation, we formulate the *evolution equation*

$$\rho_{i,t} = \rho_{i,t-1} + \delta_t$$

where $\delta_t \sim \mathcal{N}\left(0, \Delta_{\rho_{i,t}}\right)$. Thus the evolution equation is a simple re-expression of the random walk process described above. With these in place, we can simulate the posterior distributions of the IRT Responsibility scores. This occurs by first sampling $\boldsymbol{\rho}_T$, where T is the total number of time periods, from $p\left(\boldsymbol{\rho}_T \mid D_T\right)$, where D_T is shorthand for *all* information available up to time T, and then by exploiting the fact that

$$p\left(\boldsymbol{\rho} \mid D_T\right) = p\left(\boldsymbol{\rho_T} \mid D_T\right) p\left(\boldsymbol{\rho}_{T-1} \mid \boldsymbol{\rho}_T, D_{T-1}\right) \cdots p\left(\boldsymbol{\rho}_0 \mid \boldsymbol{\rho}_1, D_0\right)$$

as specified by the evolution equation. The "backward sampling" technique thus begins from the final time point and works backward to the first time point. The algorithm proceeds by computing:[1]

(a) The prior mean of $\boldsymbol{\rho}_t$ given D_{t-1},

$$\boldsymbol{a}_t = \mathbf{1}_t \boldsymbol{m}_{t-1},$$

where in the first iteration $\boldsymbol{m}_{t-1}$ comes directly from the prior on $\boldsymbol{\rho}_0$ and in subsequent iterations it has already been computed;

(b) The prior variance of $\boldsymbol{\rho}_t$ given D_{t-1},

$$\boldsymbol{R}_t = \mathbf{1}_t C_{t-1} \mathbf{1}_t' + \Delta_t,$$

where C again comes from the prior in the first iteration;

[1] For simplicity but explicitness, let $\mathbf{1}_t$ be a vector of ones with as many elements as there are observables in time period t. This is the vector to which any further covariates could be added in future applications of the algorithm.

(c) The mean of the forecast:

$$\boldsymbol{f}_t = \boldsymbol{\beta}_t \boldsymbol{a}_t$$

where we have already calculated $\boldsymbol{a}_t$ and the form comes from the evolution equation;

(d) The variance of the forecast:

$$\boldsymbol{Q}_t = \boldsymbol{\beta}_t \boldsymbol{R}_t \boldsymbol{\beta}_t' + \boldsymbol{I}_{J_t},$$

where $|J_t|$ is the number of policies in time period t;

(e) The posterior mean of $\boldsymbol{\rho}_t$ given D,

$$\boldsymbol{m}_t = \boldsymbol{a}_t + \boldsymbol{R}_t \boldsymbol{\beta}_t \boldsymbol{Q}_t^{-1} \left(\boldsymbol{z}_{t,\cdot,x} - \boldsymbol{\alpha}_t - \boldsymbol{f}_t \right)$$

where we maintain notation $(\boldsymbol{z}_{i,\cdot,t} - \boldsymbol{\alpha}_t - \boldsymbol{f}_t)$ to highlight that it is the error of the forecast; and

(f) The posterior variance of $\boldsymbol{\rho}_t$ given D_t,

$$\boldsymbol{C}_t = \boldsymbol{R}_t - \boldsymbol{R}_t \boldsymbol{\beta}_t \boldsymbol{Q}_t^{-1} \boldsymbol{Q}_t \left(\boldsymbol{R}_t \boldsymbol{\beta}_t \boldsymbol{Q}_t^{-1} \right)'.$$

This process is repeated up to T, thus providing the "forward filtering" component to the algorithm. The quantities are saved and then are utilized for the backward sampling:

(a) Sample $\boldsymbol{\rho}_T$ from a multivariate normal distribution with mean $\boldsymbol{m}_T$ and variance $\boldsymbol{C}_T$ as computed above.

(b) Calculate the mean of the conditional distribution of $\boldsymbol{\rho}_t$ given ρ_{t+1} and D_t,

$$\boldsymbol{h}_t = \boldsymbol{m}_t + \boldsymbol{C}_t \mathbf{1}_{t+1}' \boldsymbol{R}_{t+1}^{-1} \left(\boldsymbol{\rho}_{t+1} + \boldsymbol{a}_{t+1} \right)$$

where all quantities have again been computed in the forward filtering step.

(c) Calculate the variance of the conditional distribution of $\boldsymbol{\rho}_t$ given ρ_{t+1} and D_t,

$$\boldsymbol{H}_t = \boldsymbol{C}_t - \boldsymbol{C}_t \mathbf{1}'_{t+1} \boldsymbol{R}^{-1}_{t+1} \boldsymbol{R}_{t+1} \left(\boldsymbol{C}_t \mathbf{1}'_{t+1} \boldsymbol{R}^{-1}_{t+1} \right)'$$

which again are all known.

With these quantities in place, we draw from $p\left(\boldsymbol{\rho}_t \mid \boldsymbol{\rho}_{t+1}, D_t\right) \sim \mathcal{N}\left(\boldsymbol{h}_t, \boldsymbol{H}_t\right)$.

The entire process—forward filter to compute the quantities for the conditional distributions, then backward sample from them—is repeated many times. The algorithm is automated in `MCMCpack` for users in the R statistical computing environment. Importantly, given our massive data and the computational expense of the model, the automated algorithm makes use of parallel processing in C++ to maximize efficiency.

Given that we are working with such large data, and given that so many firms have adopted so few KLD items, the routine as written in `MCMCpack` only works if appropriate care is given to setting starting values. Throughout our estimation, we use the KLD Index divided by the number of metrics in year t as the starting values for $\boldsymbol{\rho}$; taking a similar approach, we use the mean number of firms that adopt a given policy as the starting values for $\boldsymbol{\alpha}$. We set the starting value of all the $\boldsymbol{\beta}$ terms at one. The model requires two firms to be "identifiers" of the dimension. We chose firms that consistently ranked high and low on routines estimated with different identifiers: IBM on the positive side and Bed, Bath, and Beyond on the negative side. Much like estimates from logit models where the variance is assumed, the parameters are estimated given assumed default settings: prior means $\boldsymbol{a}_0 = \boldsymbol{b}_0 = m_0 = 0$ and prior variance diagonal elements of $\boldsymbol{A}_0$, $\boldsymbol{B}_0$, and $\boldsymbol{C}_0$ of 0.1. We run the algorithm for 5,000 iterations after a burn-in period of 1,000 iterations; because draws may be autocorrelated, we save only every other iteration, leaving 2,500 draws from each posterior distribution of interest.

References

Albert JH, Chib S. 1993. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, **88**(422):669–679.

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.