Received: 10 June 2020 | Revised: 5 October 2021 | Accepted: 7 December 2021 | Published on: 10 January 2022

DOI: 10.1002/smj.3374

RESEARCH ARTICLE

STRATEGIC MANAGEMENT JOURNAL WILEY

Building knowledge by mapping model uncertainty in six studies of social and financial performance

Luca Berchicci[1] | **Andrew A. King**[2]

[1]Rotterdam School of Management (RSM), Erasmus University Rotterdam, Rotterdam, The Netherlands

[2]Questrom School of Business, Boston University, Boston, Massachusetts, USA

Correspondence
Luca Berchicci, Rotterdam School of Management (RSM), Erasmus University Rotterdam, Rotterdam, The Netherlands.
Email: lberchicci@rsm.nl

Abstract

Research Summary: Many scholars bemoan the difficulty of learning from individual research reports. Replication is often prescribed as a salve, but few replications are conducted, and even fewer allow the formation of a coherent understanding. In this article, we propose a complement to replication that emphasizes the mapping of epistemic uncertainties. We demonstrate our approach by exploring the results of six related studies on the link between social and financial performance. We show that our method allows the synthesis of seemingly conflicting findings, and we propose that it should be used proactively, prior to replication, to speed the growth of knowledge.

Managerial Summary: Any single empirical study provides a weak basis for inference. As a result, scholars advocate repeated analysis of important issues, but evidence from replications can be hard to integrate into a coherent understanding. For example, six important studies of the link between corporate social and financial performance have been published in this journal, but their conflicting results have defied integration. We show that a new approach to empirical research allows their reconciliation: all six suggest that across firms, social and financial performance are correlated

but that improvements in social performance seldom precede increased financial performance.

KEYWORDS

epistemology, model selection, model uncertainty, research methods, social and financial performance

1 | INTRODUCTION

To build our understanding of strategic management, or indeed any topic, we must be able to learn from the research of other scholars. Yet, the process of building knowledge from empirical testimony is fraught with uncertainty and peril. No matter how detailed the report, readers (or hearers) can seldom fully observe the empirical assumptions that gave rise to reported estimates, and thus must guess how these estimates should be interpreted (Leamer, 1985; Longino, 1990). In response to this problem, leading scholars have recommended that empirical estimates and interpretations be confirmed via replication, but practical and theoretical barriers impede the application of this strategy (Bettis, Ethiraj, Gambardella, Helfat, & Mitchell, 2016a; Bettis, Helfat, & Shaver, 2016b; Longino, 2019). Few replications are performed, even fewer are accepted for publication, and for a common kind of replication (a non-exact or "quasi" replication),[1] it can be difficult to interpret and synthesize the resulting findings (Leamer, 1985; Wilholt, 2013). In this article, we demonstrate a new method for evaluating and integrating multiple quasi-replications, and we propose that the method could also be used proactively, prior to replication, to increase the transparency of research results. In total, we demonstrate a model of research that could aid the accumulation of knowledge about strategic management.

The difficulty of synthesizing varying estimates from quasi-replications is well illustrated by six studies of the link between a firm's social and financial performance (the "SP–FP link"), and we use these studies to demonstrate the proposed method.[2] All six use the same population for analysis, similar measures, and methods, and all were accepted for publication in the *Strategic Management Journal* (Barnett & Salomon, 2012; Hillman & Keim, 2001; Hull & Rothenberg, 2008; McWilliams & Siegel, 2000; Waddock & Graves, 1997; Zhao & Murrell, 2016). Yet, these studies report divergent estimates and encourage widely differing inferences, respectively, the relationship is positive and linear, spurious, limited to particular scales, moderated, U-shaped, or conditional on certain financial measures (see Appendix S1 for more details). Given the diversity of such findings, how can practitioners and scholars build cumulative understanding?

One possibility is to focus attention on only those results reported in relatively recent publications, particularly those using more advanced empirical methods. For example, with respect to research on the SP–FP link, scholars could limit their attention to recent work using discontinuities, exogenous treatments, and instrumental variables to identify the causal connection

[1]Bettis, Helfat, and Shaver (2016b) define a "quasi replication" as one that differs from the original by using a different population (e.g., subjects or time period) or empirical procedure (e.g., measures or models).

[2]Terminology for "social performance" remains unsettled. Some scholars use "sustainability" performance or "environmental, social, governance" performance. We follow Waddock and Graves' nomenclature (Waddock & Graves, 1997), as did most of the six studies here considered.

between aspects of corporate social and financial performance (Albuquerque, Koskinen, & Zhang, 2019; Awaysheh, Heron, Perry, & Wilson, 2020; Flammer, 2015). Yet, newer results are not always more informative, because they may not be precisely on point, or because more sophisticated approaches may require sacrifices on other dimensions, such as construct precision or the degree to which the analyzed sample reflects the broader population (Leamer, 2010). In practice, older and simpler studies of associations between variables often long influence scholarship, and such is the case here: the six studies on the SP–FP link continue to be widely cited and emulated (e.g., Awaysheh et al., 2020). Yet, no consensus on their inference is in sight: some scholars accept one result as definitive, others infer contingencies, and still, others contend that no result can be trusted (Aguinis & Glavas, 2012; Orlitzky, 2011).

Another possible approach to synthesizing cumulative understanding is to use tools from meta-analysis to integrate results from a large number of studies into pooled coefficient estimates. This approach can be highly effective if based on studies using a consistent set of measures and models (Møller, Ioannidis, & Darmon, 2018). But, in emerging fields of inquiry, empirical studies can be so heterogeneous in their design that meta-analysis estimates are made unreliable (Chevret, Ferguson, & Bellomo, 2018). In the case of the SP–FP association, for example, three influential meta-analytical studies deliver conflicting conclusions (Horváthová, 2010; Margolis & Walsh, 2003; Orlitzky, Schmidt, & Rynes, 2003). In fact, the lead author of one of these studies, Marc Orlitzky, has since critiqued the use of meta-analysis, arguing that due to the varying assumptions used in research on the SP–FP link, any meta-analysis should be interpreted with great caution (Orlitzky, 2011, 2013).

In this article, we introduce another approach to synthesizing and communicating empirical research. This approach draws attention to the assumptions scholars must make to allow the calculation of coefficient estimates, and it proposes that justified inference requires transparency about the connection between empirical assumptions and estimates. We identify a space of empirical assumptions used by the original authors of the six studies and compute estimates from the implied analytical models. We show that when a few key empirical assumptions are considered, the evidence from the studies suggests a common inference: firms with higher social performance tend to have higher financial performance, yet marginal increases in social performance tend to be associated with lower, not higher, financial performance. Finally, we contend our method can be used proactively, without waiting for studies to be replicated, and thus speed the development of cumulative knowledge.

2 | THEORETICAL BACKGROUND

2.1 | The epistemology of learning from testimony about empirical research

Scholars have long understood that inference from empirical research requires assumptions (Hume, 2000). For some kinds of analysis, only a few assumptions are needed, but in many areas of social science, dozens are required (Longino, 2019). Some assumptions are based on well-established logic and evidence, others are mere guesses; some are transparently conveyed, others go unreported. Uncertain and unreported assumptions make it difficult for readers of empirical research to use published reports as a basis for justified belief (i.e., knowledge; Longino, 1990). Indeed, Gelman and Loken (2013) argue that most empirical studies require so many assumptions that the empirical process can be likened to a walk through a "garden of forking paths." At

each fork, researchers must make a choice which way to turn, and together these decisions determine where they exit the garden (i.e., the coefficients they estimate). Classical frequentist statistics are particularly vulnerable to this problem of information asymmetry because their interpretation is conditional on the use, in future samples, of an identical method (Hacking, 2001, 2016).

Faced with this problem of information asymmetry, many scholars simply adopt a position of trust with respect to published research, at least for reports that have been peer-reviewed (Fricker, 2002). However, philosophers of science have argued that simple trust provides an insufficient basis for forming justified belief (Longino, 1990, 2019; Wilholt, 2013). Since scholars make dozens or hundreds of unobservable choices in conducting their analysis, and at each point, they are influenced by personal values, the reader can fully interpret reported estimates only if she believes that the researcher's values are the same as her own (Wilholt, 2013). Returning to the metaphor of empirical research as a walk through a garden of forking paths, Wilholt is saying that a reader can use a reported estimate only if she thinks she would have made the same choice at each fork and thus exited the garden in the same place (Wilholt, 2013).

Some scholars have proposed that quasi replications can be used to map more of the paths through the empirical garden, and thereby provide information on the robustness of results to the use of rival assumptions. Yet, how conflicting results should be interpreted remains unresolved. Many quasi-replications use the same data sources and high-level empirical designs, yet still differ in many ways, both seen and unseen by readers. They may measure variables differently, or assume different functional forms, or define samples using different processes. Readers who wish to interpret conflicting results from such studies must try to discern the cause of conflicting "findings." Are they the result of known or unknown differences in the empirical method? How would estimates have changed if different assumptions had been used?

Consider, for example, the difficulty of interpreting conflicting estimates reported in just two of our six studies on the SP–FP link. Waddock and Graves (1997) estimate a positive association between social and financial performance, but McWilliams and Siegel (2000) find no significant association when R&D spending is added to the originally specified model. Based on their finding, McWilliams and Siegel (2000) invite the reader to infer that Waddock and Graves' estimate is spurious. But is this inference valid? The data and methods in the two studies differ in many known ways, and probably many unknown ones as well. Readers cannot observe, and thus must conjecture, what would have happened had Waddock & Graves estimated the model suggested by McWilliams and Siegel. Obviously, such a conjecture is little more than a guess.

To relieve the need to make guesses about counterfactuals, some scholars have proposed the use of "epistemic" or "model" uncertainty analysis. If empirical research is a walk through a garden of forking paths, they reason, then quasi replications may end up with different uninterpretable estimates and different implications, and to understand these estimates, a map of the garden is needed (Gelman & Loken, 2013; Leamer, 2010; Wilholt, 2013).

2.2 | Model uncertainty analysis

The idea of mapping estimates from many models within an empirical "garden" dates back at least to statistician Ed Leamer's proposal for "extreme bounds analysis" (Leamer, 1982, 1985, 2010). He argued that the estimates from individual studies could not be reliably interpreted because they depended on the choices and assumptions of empirical researchers. He proposed that instead of asserting the merits of a single set of assumptions and arguing for a "unique inference," scholars should consider a large set of reasonable assumptions and form estimates

from all of the implied models (Chamberlain & Leamer, 1976; Leamer, 1982, 1985, 2010). Doing so, he contended, would allow researchers to determine the "extreme bounds" of reliable inference. Unfortunately, Leamer's proposal needed modification and improvement before it could be fruitfully employed. His approach, scholars discovered, usually led to pessimistic and unhelpful conclusions. For example, Levine and Renelt (1992) evaluated various models of economic growth and reported that all were "fragile" to reasonable changes in assumptions. Sala-i-Martin (1997) advanced the approach by rejecting the "extreme bounds" perspective. He suggested that the full distribution of estimates, and not just the bounds, should be evaluated, but he did not provide guidance on how that should be done. In 2001, Fernández, Ley, and Steel (2001a, 2001b) showed that Bayesian analysis could be used to select certain models for special attention or to combine estimates into an aggregate result. In 2016, Durlauf, Navarro, and Rivers demonstrated how a model uncertainty analysis could inform the contentious debate about the effect of laws allowing concealed-carry handguns.

Along with improved analytical techniques, scholars have had to develop ways to present and explain the analysis. Durlauf et al. (2016) provide graphs of estimates sorted by marginal coefficient estimates. They then segment these graphs by various assumptions so that users can assess estimates contingent on a particular set of empirical choices. Simonsohn, Simmons, and Nelson (2020) propose that estimates can be connected to assumptions by binary codes. King, Goldfarb, and Simcoe (2021) note that in a model uncertainty analysis, the researcher no longer has the exclusive role of determining the best model and forming the best inference. As a result, they suggest that the data be presented in a manner that is useful to readers with strong or weak priors[3] about which assumptions are best. For readers with strong priors, maps between assumptions and estimates should be reported. For readers with weak priors, Bayesian analysis should be employed to direct attention to particular estimates or to form an aggregate estimate.

In this article, we use model uncertainty analysis to evaluate the link between social and financial performance. We begin by identifying the space of assumptions implied by previous research. We then estimate all of the implied models and display them graphically. For those readers with diffuse priors, we use Bayesian analysis to select a "best" model and calculate a probability-weighted combined estimate. We then interpret the totality of evidence with respect to the hypothesized link between social and financial performance, and we consider how model uncertainty analysis could be used to advance our understanding of strategic management.

3 | MODEL UNCERTAINTY ANALYSIS OF SIX QUASI-REPLICATIONS

The first step in a model uncertainty analysis is the selection of a set of assumptions to include in the assessment. Madigan and Raftery (1994) have used the metaphor of a window to explain the goal of the process. The area of the window should include those assumptions that are reasonable enough that the researcher or reader will be informed by viewing estimates obtained from models using these assumptions, and assumptions unlikely to match the true data generation process should be excluded (Madigan & Raftery, 1994). They term this set of assumptions "Occam's Window" to reflect the idea of a tradeoff between completeness and interpretability.

[3]In Bayesian statistics, *strong priors* refer to the existence of certain and unshaken beliefs on the "right" set of assumptions. *Diffuse priors* instead define beliefs that are uncertain about the right assumptions prior to observing evidence.

In practice, defining this window requires considering which empirical assumptions are justified by theory or evidence and which require guesswork. The former can be held fixed across all models in the analysis, while the latter should be allowed to vary. For example, it is well established that under certain conditions linear regression provides an efficient and unbiased way to estimate coefficients. Thus, if these conditions are met, linear regression should be specified. In contrast, it is not established how best to form scales of corporate sustainability or select control variables, and assumptions about these elements should be allowed to vary.

In cases such as ours, where the goal is to synthesize a number of quasi-replications, the assumptions used in a coherent set of focal studies can help define "Occam's window" (Durlauf et al. 2016). We searched the Clarivate Analytics Web-of-Science database to create a superset of empirical studies, in top management journals, of the SP–FP link. We then narrowed our assemblage by selecting those that were quasi-replications of the seminal study by Waddock and Graves (1997), and which did not require access to proprietary data. This left the six studies: Waddock and Graves (1997), McWilliams and Siegel (2000), Hillman and Keim (2001), Hull and Rothenberg (2008), Barnett and Salomon (2012), and Zhao and Murrell (2016).[4]

3.1 | The window of model uncertainty

Across the six studies, five modeling differences are most prominent: these involve (a) the measurement of outcome variables, (b) measurement of predictor variables, (c) functional relationships, (d) level of informative variance, and (e) appropriate controls (Figure 1). Below, we provide more detail on how we created our window of model uncertainty.

3.1.1 | Assumptions about the outcome variable

The six studies differ in their assumptions about how financial performance should be measured. Waddock and Graves (1997) assume that accounting estimates of financial performance (e.g., return on assets) best capture the construct "financial performance," but Hillman and Keim (2001) choose to use a measure of "market value-added" because they assume social performance will influence long-term firm value. Barnett and Salomon (2012) assume accounting measures are more informative about year-to-year changes, but other scholars conclude neither choice is unjustified and evaluate both market and accounting measures.

3.1.2 | Assumptions about the construction of predictor variables

Scholars also differ in how independent variables should be measured. All user data from Kinder, Lydenberg, and Domini (KLD), but they aggregate these scores in different ways. Waddock and Graves (1997) use experts to rate the importance of KLD elements and form an accordingly weighted scale. Hull and Rothenberg (2008) assume that KLD topics with more measures are more important and grant these a greater weight. Hillman and Keim (2001) argue that a subset of measures should be divided into form two estimates: "stakeholder

[4]After we completed our analysis, another paper appeared that met our criteria for selection (Awaysheh et al., 2020). We hope to include their instrumental variable approach in a future analysis.

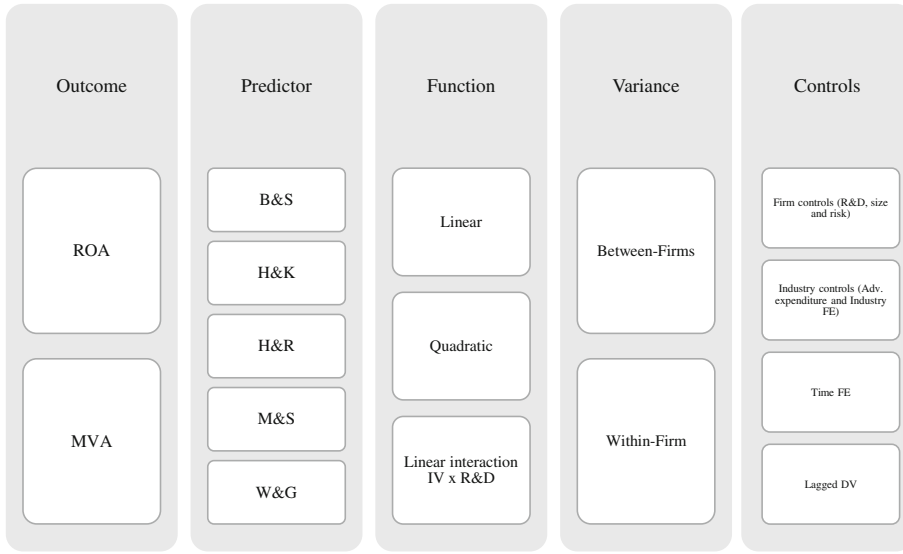


FIGURE 1 Critical assumptions in the six *SMJ* papers that determine the model space to be mapped

management" and "social issue participation." Barnett and Salomon (2012) believe that no weighting is justified and thus choose to use unitary weights for all of the KLD measures. McWilliams and Siegel (2000) argue for the use of a binary variable capturing those firms with KLD scores high enough to allow inclusion in the Domini Social Index.

3.1.3 | Assumptions about functional relationships

The scholars also differ in their assumptions about the functional form of the true relationship between social and financial performance. Most assume the relationship to be linear, but Barnett and Salomon (2012) presume that the association will be curvilinear, and they specify a model with a quadratic form. Hull and Rothenberg (2008) assume that the relationship is moderated by firm and industry attributes, and specify a model that includes interaction terms for both the firm's R&D intensity and the industry's advertising intensity.

3.1.4 | Assumptions about the level of informative variance

The six papers use two different types of variance to form estimates of the link between social and financial performance. Four analyze differences *between* firms at a cross-section in time, while two evaluate the differences *within* firms over time. Those researchers using cross-sectional analysis explicitly or implicitly assume that the mechanism for the SP–FP link will result in differences *between* firms. For example, Waddock and Graves (1997) state that they believe "good management" leads to superior social and financial performance, so the two will be correlated at a point in time. In contrast, researchers using variance *within* firms (and a lag structure) are explicitly or implicitly assuming that improved social performance should precede improved financial performance. For example, Barnett and Salomon (2012) hypothesize that firm actions concerning social performance are observed by stakeholders and then influence

their support of the firm. If so, they hypothesize, deviations from the firm's average social performance should correlate with lagged deviations from its average financial performance.

3.1.5 | Assumptions about appropriate control variables

The authors also make different assumptions about the need for control variables in their analysis. Waddock and Graves (1997) assume that some firm attributes might influence both predictor and outcome variables. As a result, they specify models that include a set of controls, usually firm size and risk (debt/assets). McWilliams and Siegel (2000) assume that both R&D and advertising intensity should be included as control variables, because they contend that these attributes are known to influence financial performance, and also may be associated with social performance. Authors also make different assumptions about the need for year fixed-effects, industry dummies, and lagged dependent variables.

3.1.6 | Sample time period

All of the authors conduct their analyses over different time periods. These differences may reflect data availability or differing assumptions about time-varying impact. Because these assumptions are not well documented in the six papers, we chose to substitute our own assumptions. We assume that the SP–FP relationship should be evident across the entire panel of data and thus choose to use all of the currently available data in our analysis. Adding time periods to our assumptions space would cause the multiplication of the number of models. For simplicity, we perform and describe our analysis using a one-time frame. However, in Appendix S3, we relax this assumption and show estimates for different time periods.

3.1.7 | The combined assumptions and implied model space

Figure 1 shows the different assumptions that determined the set of models we must estimate to form an initial map of our part of the empirical "garden of forking paths." It is comprised of two outcome variables, five alternative ways to measure the predictor variable, three functional assumptions, two types of informative variance, and 96 alternative configurations of control variables. A simple product of these options would imply 5,760 possible models, but some of these would be duplicates, given their redundant elements (e.g., including both industry fixed effects and industry advertising intensity). When we remove redundant specifications, we have 3,680 distinct models of the relationship between social and financial performance (see Appendix 2 for the computation of the final model set).

3.2 | Analytical process

3.2.1 | Data sources and sample creation

The social investment research firm Kinder, Lydenberg, and Domini (KLD) began reporting data on firm social and environmental dimensions in 1991. Over time, KLD data became a

common choice for research on the SP–FP link. Dozens of papers have been published using KLD data, including the six we consider in this analysis.

In 1991, KLD reported an evaluation of an initial sample of 650 firms (mainly S&P 500 firms). In 2001, it increased the sample to include the 1,000 largest firms in the United States. In 2003, KLD expanded that coverage to the largest 3,000 US companies by market capitalization. And, in 2013, KLD expanded its analysis to non-US companies.

Over the years, KLD measures have expanded as well. The measure started with eight topic areas, five of which (employee relations, community relations, product, environment, and diversity) were measured both as strengths and weaknesses. Currently, there are 13 topic areas in the KLD data: for seven, firms receive scores for both strengths and weaknesses, and for six, firms receive only weaknesses (e.g., alcohol, gambling, tobacco, and firearms).

To measure each firm's financial performance, we obtained financial data from Compustat and CRSP at Wharton's WRDS data center. These data provide us with all of our firm-level financial measures, along with the counts of employees, sales, and industry affiliation, and industry level data such as advertising intensity.

To combine the two data sets, we matched different types of information from the KLD and Compustat data. KLD data includes name and stock ticker and some CUSIP numbers. Compustat data includes name and CUSIP numbers and some ticker information. Unfortunately, the two systems differ in how they update information as it changes over time: Compustat backdates all changes, while KLD does not. In addition, IDs are not always recorded consistently. Thus, to match the two sets of data, we constructed a program that employed various alternatives for name and alternative IDs. Matches were ranked and then checked for accuracy by the authors. In total, we were able to match 5,986 firms for 42,736 firm-year observations between 1991 and 2015.

3.2.2 | Restricted and unrestricted samples

As with every dataset, ours has several issues that require researcher judgment. To allow visual and Bayesian comparison of models, we need to use a consistent sample. The need for such consistency implies that we must remove observations from the analysis if the full set of values is not present. Measures of R&D were found to be most frequently missing (20,770 times in all).

In very few cases, we also chose to eliminate observations with reported measures many standard deviations outside the norm. Specifically, our measure of return on assets contains eight observations with values below negative 500% or above positive 500%. Given the median ROA is 3.48% in the full sample, we decided to exclude those observations. Likewise, we decided not to include six observations where reported R&D spending exceeded revenue.

Our final dataset includes 15,061 observations covering 2,562 firms, the average of which is observed over 7 years. To make sure our restrictions did not bias our results, we also tested the reliability of our results by graphing estimates from the full unrestricted sample.

3.2.3 | Variable construction

Dependent variables

As discussed earlier, the six studies used a range of accounting and market-based firm performance measurements. We select the most common form of each type to analyze. We calculate

Return on Assets (ROA) as net income divided by the total assets of the firm in a given year. We then multiply this raw value by 100 to obtain the percentage transformation. To calculate the Market to Book Ratio, we calculate the asset value of all shares divided by the sum of all book assets and liabilities.

Independent variables

The six studies use five scales of social performance, and we replicate each of these scales. The first, *W&G*, follows the method proposed by Waddock and Graves (1997). Their scale is based on expert opinions about the importance of eight KLD social rating dimensions (i.e., employee relations, product, community relations, environment, treatment of women and minorities, nuclear power, military contracts, and South Africa).

The Hull and Rothenberg (2008) scale uses a different giving "a disproportionally greater weight than those with only one subcategory" (Hull & Rothenberg, 2008, p. 784). Unfortunately, they do not specify which categories they use or how much weight is given for each subcategory. We use the main categories and assume linear weights to form our scale *H&R*.

A third scale, *H&K* is proposed by Hillman and Keim (2001). It is the sum of the strengths minus the sum of the weaknesses of five social rating dimensions (employee relations, diversity issues, product issues, community relations, and environmental issues). For the sake of parsimony and ease of comparison, we do not include their measure of social issue participation.

A fourth scale, *B&S* is based on a scale proposed by Barnett and Salomon (2012). As with the previous variables, it aggregates strengths minus weaknesses. Unlike the previous scales, it includes all 13 KLD social performance criteria.

Finally, *M&S*, is based on the measure proposed by McWilliams and Siegel (2000) who create a dummy variable indicating the inclusion of the firm in Kinder, Lydenberg, and Domini's index of 400 firms with top sustainability performance. The index includes companies with the highest KLD ratings in each sector, but also eliminates some companies based on values screens: alcohol, tobacco, gambling, and so on. We build our version of the M&S measure by creating a dummy variable equal to 1 if a firm is in the index in that year and zero otherwise.

To allow the comparison of effect sizes, we normalize all of the continuous scales so that they have a unitary standard deviation.

Moderating and mediating variables

Both Hull and Rothenberg (2008) and McWilliams and Siegel (2000) argue that R&D spending moderates or mediates the relationship between social and financial performance. To create our measure of R&D, we divide the firm's annual R&D spending by its total sales in each year.

Control variables

Following prior research, we calculate the firm size using sales ($M), risk as to the firm debt/asset ratio, and advertising intensity (*adv*) as advertising expenditure divided by sales. We also calculate two sets of dummy variables to capture *industry* effects at the SIC two-digit level and *year* effects. Finally, we follow Barnett and Salomon (2012) by calculating a lagged dependent variable as FP_{t-1}.

Table 1 shows the descriptive statistics for the restricted sample of 15,061 observations. Given the effort the original authors placed on the creation of novel scales, we were surprised to find that all the four SP continuous independent variables were highly correlated, ranging from 0.95 to 0.98. Even the binary measure, M&S, is correlated with the others at 0.56 or higher.

TABLE 1 Descriptive statistics

Variable	Mean	*SD*	Min	Max									
ROA	4.240	12.128	−317.31	69.50	1								
MTB	2.487	4.276	−247.96	139.61	0.112	1							
B&S	−0.126	1.102	−5.22	7.23	0.072	0.015	1						
H&K	0.055	1.123	−4.41	7.50	0.084	0.013	0.950	1					
H&R	−0.057	1.110	−4.55	7.04	0.08	0.014	0.985	0.968	1				
M&S	0.241	0.428	0	1	0.104	0.035	0.566	0.580	0.565	1			
W&G	0.025	1.120	−4.51	7.29	0.083	0.013	0.958	0.995	0.965	0.578	1		
R&D	0.068	0.106	0	0.99	−0.324	0.015	0.053	0.05	0.037	−0.039	0.058	1	
Size	6,089.423	21,733.860	0.195	483,521	0.058	0	−0.007	0.078	0.040	0.082	0.059	−0.091	1
Risk	0.187	0.197	0	2.10	−0.15	0.019	−0.054	−0.044	−0.048	−0.063	−0.045	−0.19	0.007	1
Adv	4.363	1.344	0.19	7.66	0.052	0.011	0.031	0.066	0.055	0.149	0.054	0.035	0.115	−0.062

Note: Number of observations: 15,061.

3.2.4 | Statistical analysis

The six studies imply models using either the between-firm variance or within-firm variance. We estimated the former using a "between" estimator in which firm-averaged measures of independent variables are regressed on firm-averaged measures of the dependent variable.

$$\bar{y}_i = \alpha + \bar{x}_i\beta + \nu_i + \epsilon_i \quad (1)$$

Where $\nu_i \& \epsilon_i$ are the firm-specific and iid error terms. Note that a critical assumption required for this analysis is that estimates of β will be unbiased by the correlation between ν_i and $\bar{x}_i$.

We estimate models using within-firm variance by specifying a fixed effect for each firm. This means that all measures are demeaned by firm averages.

$$(y_{it+1} - \bar{y}_i) = (x_{it} - \bar{x}_i)\beta + (\epsilon_{it} - \bar{\epsilon}_i) \quad (2)$$

We estimate coefficients for all 3,680 models and store the estimate of the SP coefficient at the mean, its standard error, its confidence interval, and the R^2 of the estimated model. We also store the log-likelihood of each model.

3.2.5 | Reporting a model uncertainty analysis

Model uncertainty analysis shifts the role of the researcher from authority to coach. The researcher no longer has the task of uncovering and reporting "the unique inference [that] can be squeezed from a data set" (Leamer, 1985, p. 308). Instead, she is now responsible for presenting her empirical analysis in a manner that allows readers to form their own inferences. As Leamer (1985) notes, this implies a switch from the perspective of a classical statistician to that of a Bayesian: inferences become contingent on the beliefs that readers bring to the analysis.[5]

King et al. (2021) suggest that estimates should be presented so that readers are encouraged to consider their prior assumptions in forming inferences. Durlauf, Navarro, and Rivers (2016) propose that estimates be sorted by those assumptions about which readers might have differing opinions. They also suggest reporting confidence intervals so that readers can gauge the reliability of a particular model, but these must be interpreted in a strict Neyman/Pearson manner. That is, the interval calculated by the employed method has a set probability (usually 95%) of producing an interval that includes the true population parameter.

Durlauf et al. (2016) also suggest that the full set of estimates be analyzed in a strictly Bayesian manner. Informally, this requires discussing any inference as conditional on prior assumptions. Formally, it allows the calculation of posterior probabilities conditional on a set of prior beliefs, and for simplicity, diffuse priors are assumed. In plain English, such a Bayesian analysis allows the identification of more probable models, but only for those who initially believe that every model has an equal probability of matching the true data-generation process.

Graphing

After running all of the models and storing their estimates, we used the STATA graphic interface to display the marginal effects of the coefficient of "social performance" (SP). For Figures 2–

[5]Frequentist tests require a prespecified sampling and analysis plan.

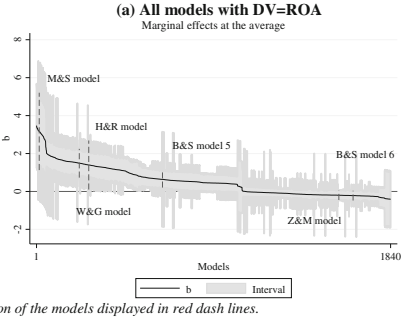
4, we show the average marginal effect of the predictor estimated at the mean of social performance (labeled as "b"), as well as the lower and upper bound of the 95% confidence interval for this estimate. To allow reader inference, we apply for different sorting orders before graphing. In addition, we mark models that are closest to those used in the six papers in Appendix S1 with red lines capturing the 95% confidence interval for our estimate. In Figure 5a,b, we graph the average marginal effect of the coefficient of the interaction term (SP * R&D), while in Figure 5c, d, we split the sample at the median of R&D to take into account the within-firm variance (Shaver, 2019). Thus, Figure 5c,d show the marginal effects of b(SP) above and below the median of R&D. In Figure 6a,b, we show the effect of any curvilinear relationship by graphing both the marginal effect of social performance, $b(x)$, and the marginal effect of its squared term, $b(x^2)$ for the models with between-firm variance. For the within-firm variance models, in Figure 6c,d, we follow Shaver's approach (Shaver, 2019), and show the marginal effect of $b(x)$ by splitting the sample by the median of SP.

Bayesian model selection and averaging

Bayesian analysis provides a way to synthesize the numerous estimates made in a model uncertainty analysis. Bayes' law quantifies how a rational actor should update her beliefs based on new information: the probable truth of a hypothesis H_i after seeing data is a function of the probability that such data would be observed if H_i was indeed true, $P(D|H_i)$, the prior beliefs about the truth of that hypothesis $P(H_i)$, and the probability that the data would be observed by any means, $P(D)$, where there are n possible explanations for the data.

$$P(H_i|D) = \frac{P(D|H_i) * P(H_i)}{P(D)} = \frac{P(D|H_i) * P(H_i)}{\sum_1^n P(D|H_i) * P(H_i)} \tag{3}$$

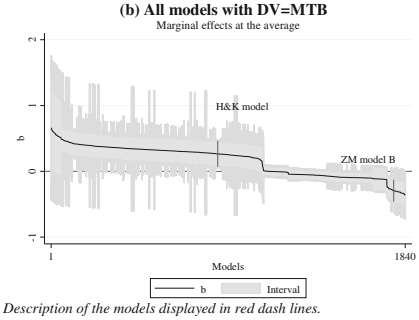
In a model uncertainty analysis, we are interested in determining the probability that a particular model matches the true data-generation process:



Description of the models displayed in red dash lines.

- "W&G model" refers to the Model 1, table 6 in Waddock & Graves (1997);
- "M&S model" refers to the Model 3, table 2 in McWilliams & Siegel (2000);
- "H&R model" to Model 3, table 2 in Hull and Rothenberg(2008);
- "B&S model 5" to Model 5, table 2 Barnett and Salomon (2012);
- "B&S model 6" to Model 6, table 2 Barnett and Salomon (2012); and
- "Z&M model" to table 4 with DV = ROA in Zhao and Murrell (2016).

Over the collection of models with ROA as DV, 59% of the coefficient estimates are positive and 41% negative. 41% of the models result an estimate of a positive coefficient with a 95% confidence interval not inclusive of zero; 2.1% of models result an estimate of a negative coefficient with a 95% confidence interval not inclusive of zero.

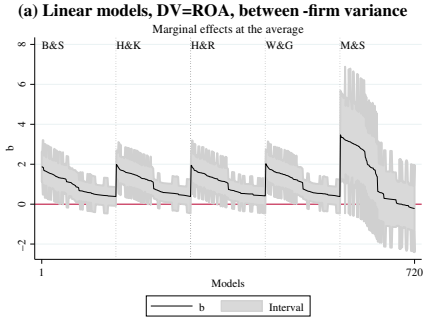


Description of the models displayed in red dash lines.

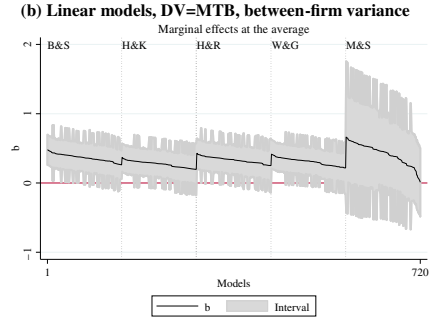
- "H&K model" to table 2 in Hillman and Klein (2001);
- "Z&M model B" to Table 4 with DV = MTB in Zhao and Murrell (2016).

Over the collection of models with MTB as DV, 62% of the coefficient estimates are positive and 38% negative. 41% of the models result an estimate of a positive coefficient with a 95% confidence interval not inclusive of zero; 11 % of models result an estimate of a negative coefficient with a 95% confidence interval not inclusive of zero.

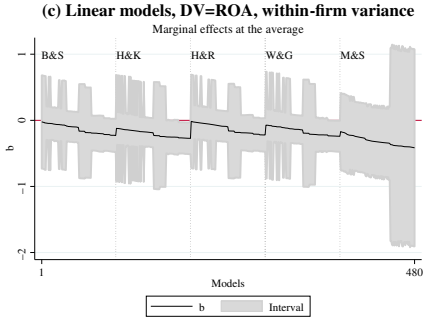
FIGURE 2 (a) All models with DV = ROA. (b) All models with DV = MTB



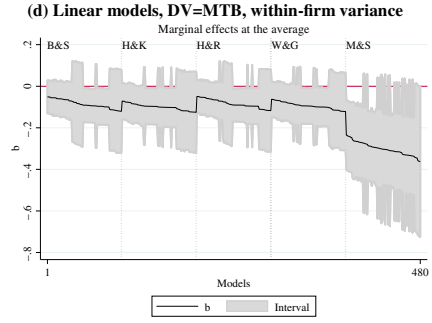
Pos.	100%	100%	100%	100%	82%
CI n/i 0*	58%	76%	68%	74%	32%
Neg.	–	–	–	–	18%
CI n/i 0*	–	–	–	–	–



Pos.	100%	100%	100%	100%	100%
CI n/i 0*	76%	65%	67%	67%	1%
Neg.	–	–	–	–	–
CI n/i 0*	–	–	–	–	–



Pos.	–	–	–	–	–
CI n/i 0*	–	–	–	–	–
Neg.	100%	100%	100%	100%	100%
CI n/i 0*	8%	33%	–	25%	–



Pos.	–	–	–	–	–
CI n/i 0*	–	–	–	–	–
Neg.	100%	100%	100%	100%	100%
CI n/i 0*	33%	42%	33%	33%	68%

FIGURE 3 (a) Linear models, DV = ROA, between-firm variance. (b) Linear models, DV = MTB, between-firm variance. (c) Linear models, DV = ROA, within-firm variance. (d) Linear models, DV = MTB, within-firm variance

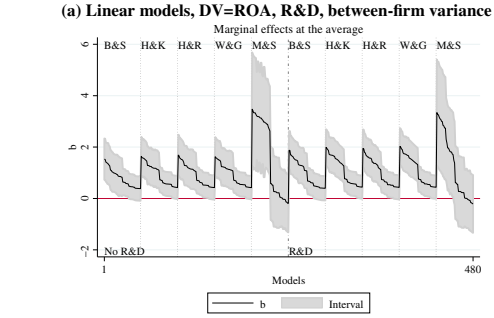
$$P(M_i|D) = \frac{P(D|M_i) * P(M_i)}{\sum_1^n P(D|M_i) * P(M_i)} \tag{4}$$

Given the number of possible ways that the data could be generated, the probability that any given model is correct, $P(M_i|D)$, is usually very small. Yet, its measure provides valuable information because it allows us to estimate the *relative* probability of a given model in a set.

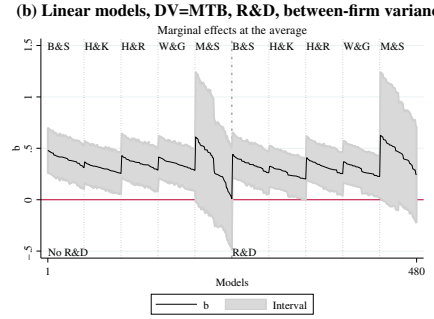
We still face the difficulty that Equation (4) includes prior probabilities for each model, $P(M_i)$, and these unobserved priors may differ across researchers and readers. However, if we assume that all models within a given window are equally likely, we can simplify the equation and eliminate the priors. If $P(M_i) = \mathrm{c}, \text{and} \sum_i^n c = \mathrm{K}$. Equation (4) can now be simplified to:

$$P(M_i|D) = \frac{P(D|M_i)}{\sum_1^n (P(D|M_i)} \tag{5}$$

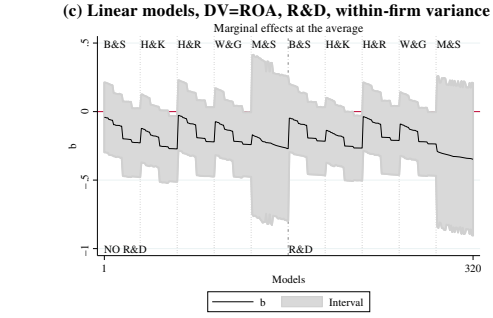
We are still faced with the calculation of $P(D|M_i)$, but fortunately, Chib (1995) showed that, due to Bayes' Rule, $P(D|M_i)$ can be expressed as:



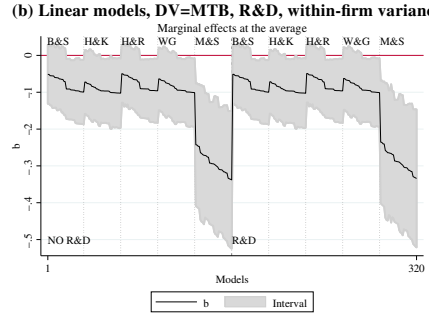
Pos.	100%	83%	100%	83%
CI n/i 0*	67%–85%	50%	67%–93%	46%
Neg.	–	16%	–	16%
CI n/i 0*	–	–	–	–



Pos.	100%	100%	100%	100%
CI n/i 0*	100%	–	95%–100%	4%
Neg.	–	–	–	–
CI n/i 0*	–	–	–	–



Pos.	–	–	–	–
CI n/i 0*	–	–	–	–
Neg.	100%	100%	100%	100%
CI n/i 0*	0%–25%	0%	0%–25%	0%



Pos.	–	–	–	–
CI n/i 0*	–	–	–	–
Neg.	100%	100%	100%	100%
CI n/i 0*	50%–62%	100%	50%–62%	100%

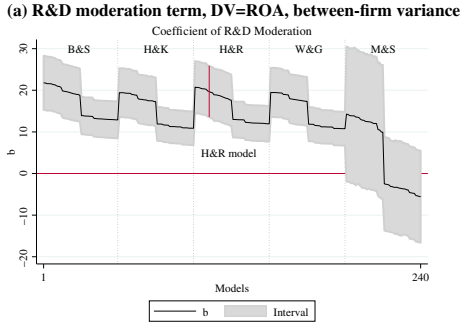
FIGURE 4 (a) Linear models, DV = ROA, R&D, between-firm variance. (b) Linear models, DV = MTB, R&D, between-firm variance. (c) Linear models, DV = ROA, R&D, within-firm variance. (d) Linear models, DV = MTB, R&D, within-firm variance

$$P(D|M_i) = \frac{P(D|M_i,\theta_{M_i})P(\theta_{M_i}|M_i)}{P(D|\mathrm{M}_i,\theta_{M_i})} \tag{6}$$

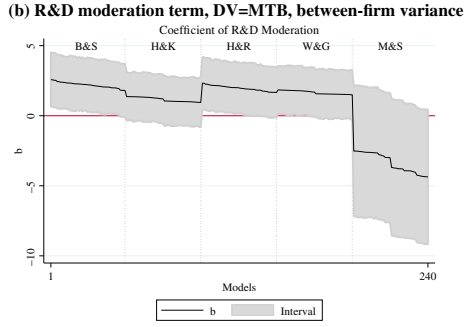
and that this relationship holds for any value of the parameters θ_{M_i}. He recommends using the parameter $(\widehat{\theta})$ at the high-density point of the posterior distribution, or the maximum likelihood, of the estimator: $\mathrm{L}\left(\widehat{\theta},M_i\right)$. For most statistical packages, this is easily obtained for all of the specifications used in our analysis. We can now calculate the likelihood for each model and select the one with the greatest likelihood as a reference model (M_r). We can then calculate the relative probability of every model relative to that reference.

$$\sum_i^n \frac{P(D\,|\,M_i)}{P(D\,|\,M_r)} = \sum_i^n \frac{L\left(\widehat{\theta},M_i\right)}{L\left(\widehat{\theta},M_r\right)} = 1 \tag{7}$$

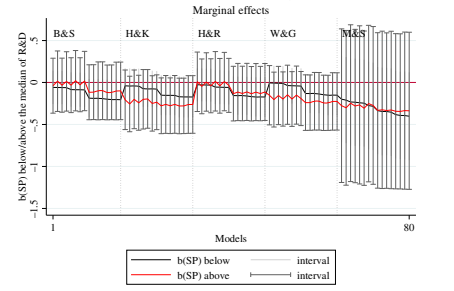
Further, we can use these relative posterior probabilities to form a weighted average coefficient.



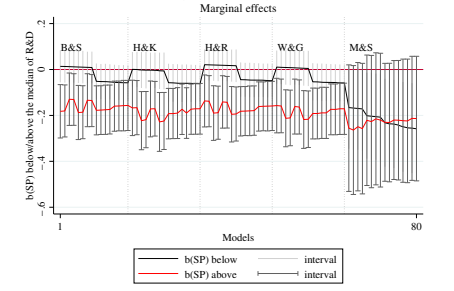
Pos.	100%	50%
CI n/i 0*	100%	–
Neg.	–	50%
CI n/i 0*	–	–



Pos.	100%	100%	100%	100%	–
CI n/i 0*	92%	0%	67%	42%	–
Neg.	–	–	–	–	100%
CI n/i 0*	–	–	–	–	–



Pos. Above	25% [–]	–	25% [–]	–	–
Neg. Above	75% [–]	100% [–]	75% [–]	100% [–]	100% [–]
Pos. Below	–	–	–	–	–
Neg. Below	100% [–]	100% [–]	100% [–]	100% [–]	100% [–]
CI n/i 0* between square brackets					



Pos. Above	–	–	–	–	–
Neg. Above	100% [100%]	100% [100%]	100% [100%]	100% [100%]	100% [–]
Pos. Below	50% [–]	12% [–]	50% [–]	50% [–]	–
Neg. Below	50% [–]	88% [–]	50% [–]	50% [–]	100% [75%]
CI n/i 0* between square brackets					

FIGURE 5 (a) R&D moderation term, DV = ROA, between-firm variance. (b) R&D moderation term, DV = MTB, between-firm variance. (c) R&D moderation term, DV = ROA, within-firm variance. (d) R&D moderation term, DV = MTB, within-firm variance

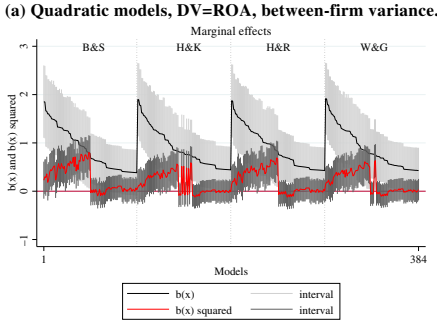
$$E\left(\overline{B}\middle| D\right)=\sum_{i}^{n}B_i\frac{L\left(\widehat{\theta},M_i\right)}{L\left(\widehat{\theta},M_r\right)} \tag{8}$$

Assuming that the estimate variance s^2 of each model is independent,[6] we can calculate the variance of this average coefficient estimate as:

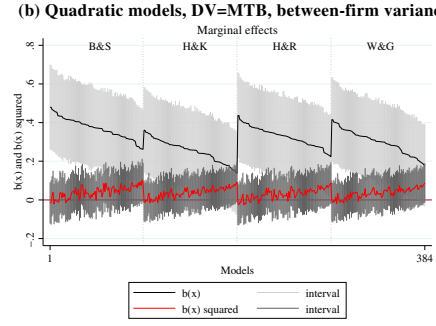
$$E\left(\overline{s}\middle| D\right)=\sqrt{\left\{\sum_{i}^{n}s_i^2\frac{L\left(\widehat{\theta},M_i\right)}{L\left(\widehat{\theta},M_r\right)}\right\}} \tag{9}$$

Finally, we can select a "best" model for special attention based on the model with the highest likelihood.

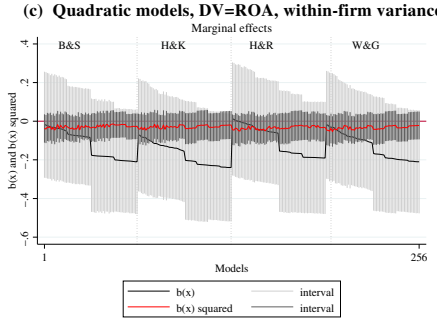
[6]In future work, we hope to use a Monte Carlo method to construct weighted sample variance.



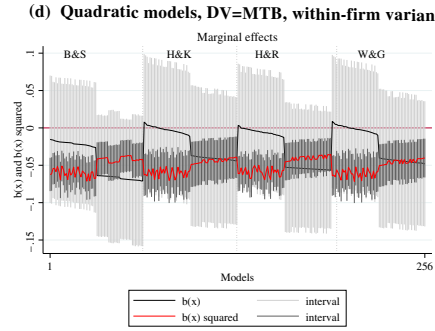
Pos. b(x)	100% [67%]	100% [81%]	100% [80%]	100% [83%]
Neg. b(x)	–	–	–	–
Pos. b(x) sq.	84% [–]	65% [24%]	78% [36%]	73% [33%]
Neg. b(x) sq.	16% [–]	35% [–]	22% [–]	27% [–]
CI n/i 0* between square brackets				



Pos. b(x)	100% [100%]	100% [74%]	100% [100%]	100% [92%]
Neg. b(x)	–	–	–	–
Pos. b(x) sq.	84% [–]	65% [24%]	78% [36%]	73% [33%]
Neg. b(x) sq.	16% [–]	35% [–]	22% [–]	27% [–]
CI n/i 0* between square brackets				



Pos. b(x)	100% [–]	–	100% [–]	–
Neg. b(x)	–	100% [–]	–	100% [–]
Pos. b(x) sq.	–	–	–	–
Neg. b(x) sq.	100% [–]	100% [52%]	100% [14%]	100% [–]
CI n/i 0* between square brackets				



Pos. b(x)	100% [50%]	100% [–]	100% [50%]	100% [–]
Neg. b(x)	–	–	–	–
Pos. b(x) sq.	–	–	–	–
Neg. b(x) sq.	100% [100%]	100% [100%]	100% [100%]	100% [100%]
CI n/i 0* between square brackets				

FIGURE 6 (a) Quadratic models, DV = ROA, between-firm variance. (b) Quadratic models, DV = MTB, between-firm variance. (c) Quadratic models, DV = ROA, within-firm variance. (d) Quadratic models, DV = MTB, within-firm variance

4 | RESULTS

We first present our analysis graphically and then report the results of our Bayesian analysis.

4.1 | Graphical analysis

Figure 2 shows our estimates for all 3,680 models, sorted by the size of the marginal relationship between SP and FP, calculated at the mean value of the predictor variables (solid line). The 95% confidence interval is displayed as a gray region. For estimates of the relationship between SP and ROA (Figure 2a), coefficient estimates range from 3.78 to −0.38. For SP and MTB (Figure 2b), the range is 0.75 to −0.20. The overall range of results provides a visual impression of the ambiguity created by epistemic (model) uncertainty and aleatoric (statistical) uncertainty. For many of the models, the confidence interval measuring aleatoric uncertainty (the gray region) is smaller than the uncertainty created by model selection (i.e., the range of estimates

for b across the models). This reinforces the importance of considering model uncertainty when forming inferences.

We indicate with red lines those estimates from models closest to those reported by the six previous authors. These estimates do not represent exact replications because our sample differs from previous studies; instead, they reflect what those authors might have found had they used *our* sample in *their* analysis. As expected, models used by Waddock and Graves (1997) and Hillman and Keim (2001) deliver estimates of a positive relationship. For our approximation of Hull and Rothenberg (2008, Model 3), we also find the expected positive main effect. Consistent with Barnett and Salomon (2012) and Zhao and Murrell (2016), we find that some of their models can deliver nonsignificant or negative estimates of the marginal relationship between SP and FP, a topic we will return to later. Strikingly, our replication of the model proposed by McWilliams and Siegel (2000) results in a very different estimate than the one they report. In contrast to their negative estimate, our "M&S" model delivers one of the strongest positive coefficients of all! However, this should not be interpreted as demonstrating their reported estimate is erroneous, because our estimate is based on a larger sample over a longer time period. We will return to their model and estimate below.

Despite verifying the sign and significance of some of the estimates in the six reports, our analysis reveals the difficulty of building knowledge from published quasi-replications. Our analysis also shows that, for the SP–FP relationship, future replications are unlikely to allow the building of shared understanding. Consider what would happen were future researchers to pick one of the 3,680 models and conduct a traditional hypothesis test. Our analysis suggests that each researcher would have a 52% chance of concluding they could not reject $B = 0$, a 41% chance of *rejecting* a hypothesis that $B < 0$, and a 6% chance of *rejecting* the hypothesis that there is a *positive* relationship between SP and FP ($B > 0$). Given the number of scholars interested in the SP–FP relationship, it seems likely that significant results, in both directions, would be found and reported. If "surprising" or disconfirming estimates were prioritized for publication, reported findings might jump back and forth as predicted by De Long and Lang (1992). In such a circumstance, we contend, consensus would be hard to obtain, as indeed has been the case in the SP–FP literature.

4.1.1 | Meaningful variance

The six teams of authors make different assumptions about what kind of variance should be used to evaluate the SP–FP relationship. Figure 3 shows the estimates for all linear models (quadratic models will be discussed later) using only between-firm variance (Figure 3a,b) or within-firm variance (Figure 3c,d). As discussed earlier, when exploiting between-firm variance, estimates are made using average firm values over the panel, and when exploiting within-firm variance, estimates are made using values demeaned by the firm's average for that variable. In Figure 3, we sort the variables so that the different scales of social performance are grouped together.

We interpret Figure 3 as making it clear that the employed level of variance (between- or within-firm) has a strong effect on the estimated relationship. When between-firm variance is used (Figure 3a,b), coefficient estimates are mostly (for ROA) or always positive (for MTB). In contrast, when within-firm variance is used (Figure 3c,d), coefficient estimates are always negative, though the confidence interval often includes zero. In its entirety, we contend that Figure 3 suggests that inference should be contingent on the researcher's (or reader's) prior beliefs about how the SP–FP relationship manifests itself.

4.1.2 | Scales of social performance

Our analysis also helps to reveal where differing assumptions *do not* lead to divergent estimates. For example, the authors of the six focal articles seemed to believe that different scales of KLD data would deliver different results. They debate scales weighted by expert ratings, the number of topics within a category, and the connection to management versus participation. In some cases, they infer that the use of different scales explains differences in the findings among the studies. Yet, as shown in Figure 3, for the four continuous scales, the pattern of results is remarkably similar. Only the binary measure proposed by McWilliams and Siegel results in a sharply different pattern of results; we conjecture that the greater range of estimates, and their higher standard errors, resulting from the loss in resolution caused by dichotomization.

4.1.3 | R&D mediation

Two of the original research projects proposed that R&D mediates or moderates the SP–FP relationship. In Figure 4, we sort and graph the results by (a) no mediation ("No R&D"), and (b) mediation only ("R&D"). (Models with R&D moderation are shown in Figure 5.) Within these categories, we continue to sort by the measure for social performance. To our eyes, the graphs make it evident that, regardless of the scale used, mediation has little effect on the estimate of the main link between SP and FP. This is true regardless of the variance used or the specification of the dependent variable. Thus, we conclude that it is rational to reduce belief in a mediating effect of R&D. In future work, we plan to explore why our estimates differ from previous research; some possible areas for investigation include differences in sample selection, matching methods, and the use, in previous work, of imputed R&D measures.

4.1.4 | R&D moderation

Figure 5a,b shows the coefficient for and R&D moderating term (not the main effect) for all linear models, including that term. Figure 5a reveals that most models using between-firm variance and ROA estimate a strong positive moderation. Note that for this case, Hull and Rothenberg (2008) found a negative relationship. The only exception is when McWilliams and Siegel's binary measure is used. We thus infer that we can confirm only part of H&R's analysis: as shown in Figure 2, we confirm that their assumptions and specifications are consistent with a main positive relationship, but in Figure 5a, we show that, across our model space, we estimate a moderation term that is reliably the opposite of the one they reported.

In Figure 5c,d, we report coefficient estimates for models using within-firm variance to evaluate R&D moderation. To show the effect of moderation, we cannot simply graph the coefficient for the moderating term, because as Shaver (2019) describes such an estimate would include both within-firm and between-firm effects. Thus, we follow best practice by instead splitting the sample by the median value of the variable of interest, and then report marginal effects for the two samples.

Figure 5c shows the marginal effect of SP on ROA for high R&D firms (red line) and low R&D firms (black line). For both groups, the estimated coefficients of SP are negative and broadly similar, and their intervals overlap extensively, suggesting little moderation. When the

outcome variable is MTB (Figure 5d), the marginal beta for high R&D firms is evidently lower than those for low R&D firms (black line), suggesting a negative moderation.

4.1.5 | Curvilinear relationship

Barnett and Salomon (2012) conjecture that the link between social and financial performance is contingent on the initial level of social performance. Those firms with low or high social performance will gain financially, they hypothesize, if the firm improves its social performance, but firms of average performance will not. B&S operationalize their conjecture by specifying a quadratic model and report mixed results. When measuring financial performance as ROA and making estimates with between-firm variance, they find a strong significant positive quadratic U-shaped relationship, but when using a within-firm variance, they find no significant evidence of a relationship.

To allow a better interpretation of the curvilinear relationship, Figures 6a,b show the marginal effects of both $b(x)$ and $b(x^2)$ coefficients for the models specifying a curvilinear form (further sorted by the SP scale and descending $b(x)$). In models with between-firm estimations, Figure 6a, the $b(x)$ coefficient alone is always positive across models; and $b(x^2)$ coefficient is positive for a 67–85% range of the cases. None of the models result in the negative estimates for $b(x)$ that are needed to support Barnett and Salomon's proposal for a "U" shaped relationship. Our models instead suggest that, conditional on some empirical assumptions, the relationship is best characterized as linear [i.e., $b(x^2)$ is small], and for other assumptions, it is best characterized as a half U—curvilinear, but always rising.

Figure 6c,d shows within-firm estimations of the proposed curvilinear relationship. Following Shaver (2019), we split the sample by the median of SP and run the models for firms with values below and above the median of SP. Figure 6c (with DV = ROA) shows that the relationship between SP and FP is best characterized as linear and convex. Figure 6d suggests that the relationship between SP and FP for models with the MTB as DV is characterized as a more extreme concave shape, downward sloping, and increasingly so. Thus, assumptions about the appropriate level of variance are again critical for the interpretation of our estimates.

4.1.6 | Summary of graphical analysis

We believe that the above graphical analysis demonstrates the importance of researcher (and reader) priors in guiding how estimates are interpreted. Imagine, for example, a researcher who believes that between-firm analysis best captures the true SP–FP data-generation process. If she were to enter this particular "garden of forking paths," she would proceed down a broad highway leading to a single destination. Regardless of her beliefs and choices with respect to other aspects of her model (scales for social performance, measurement of financial performance, R&D mediation, R&D moderation, and functional forms), she would end up estimating coefficients that allowed her to infer a positive association between social and financial performance. For many modeling choices, she would also find robust evidence for positive R&D moderation.

But suppose our imaginary traveler thought the SP–FP relationship should be estimated using within-firm variance. For most paths through the garden, she would estimate a negative coefficient for the SP–FP relationship, but the strength and aleatoric uncertainty of this coefficient would depend on her other choices. And, estimates with respect to R&D moderation or

curvilinear form would depend on a variety of other assumptions. Depending on her priors about the proper outcome variable, a researcher or reader would form very different inferences about R&D moderation or curvilinear form.

4.2 | Bayesian analysis for readers with diffuse priors

Readers that are unsure which assumptions are more appropriate can still make sense of the totality of the results by emphasizing estimates from models that are believed to be more probable *after* the data have been analyzed. As discussed earlier, conditional on a set of prior beliefs about the models, we can use Bayesian methods to identify the model that is most probable a posteriori. We can also use these posterior probabilities to calculate an aggregate estimate.[7]

In Appendix S4, we provide estimates sorted by the lagged dependent variables. Figure S4.1a shows that the positive SP coefficients of models with lagged ROA tend to be smaller and their confidence intervals are more likely to include zero than models without lagged ROA. This might indicate evidence of the problem identified by Achen (2000). This pattern is less evident in models with and without lagged MTB (Figure S4.1b).

Another limitation of the method is that we cannot make meaningful comparisons of posterior probabilities across models using different levels of variance or samples. Thus, we need to calculate likelihood ratios for a given variable of interest, variance level, and sample. As shown in Table 2, we conduct our Bayesian analysis within a 2×2 of models (variance: between or within-firm variance; outcome: ROA or MTB).

Table 2 shows the specification of the models that have the highest posterior probability for each of the four groups and the resulting estimated marginal relationship between SP on FP. If the highest probability model includes a moderating term, this coefficient is reported as well. The bottom of the table displays a posteriori probability-weighted synthesized estimate of the marginal effect (labeled Probability Weighted Estimated).

The "best" models using between-firm variance both used Barnett & Salomon's scale (Barnett & Salomon, 2012) of social performance, and those using within-firm variance both use Hillman and Keim's (2001). All of the "best" models included a lagged dependent variable and R&D. For two groups, the best model included the interaction term SP × R&D, suggesting that R&D moderation increases the model's fit with the data. Other control variables used in "best" models include sales, risk, and year dummies.

The reported estimates are consistent with our graphical analysis. For models using between-firm variance, the "best" model results in estimates of either a main positive relationship or a positive moderated one. For models using within-firm variance, the "best" model estimates a main negative relationship or a negative moderated one. Thus, a reader with diffuse priors about proper assumptions of models could conclude that inference is conditional on the

[7]A serious limitation to this approach is that it requires specification of prior probabilities. To simplify the analysis, scholars have tended to assume that all assumptions are equally likely. We follow this approach, though we note a concern raised by our reviewers: adding one of our possible controls, a lagged dependent variable, causes a violation of the conditions under which regression analysis provides the best estimator, and might cause coefficients of interest to be biased downward (Achen, 2000). Unfortunately, removing the lagged dependent variable can cause other problems, and the literature remains incomplete on whether inclusion or exclusion will provide the most accurate estimates (Keele & Kelly, 2006). Thus, we choose to include both model specifications.

TABLE 2 "Best" and probability-weighted estimates of SP–FP association

	Groups within which models are compared			
Variance	**Between**	**Between**	**Within**	**Within**
DV	$ROA_{(t+1)}$	$MTB_{(t+1)}$	$ROA_{(t+1)}$	$MTB_{(t+1)}$
	Specification with highest posterior probability (P[M|E])			
SP scale	B&S	B&S	H&K	H&K
DV(t)	Yes	Yes	Yes	Yes
R&D	Yes	Yes	Yes	Yes
SP × R&D	Yes			Yes
Risk		Yes		Yes
Sales				Yes
Year FE			Yes	Yes
Relative $P(M\|D)$	53%	30%	42%	89%
	Marginal effect at mean for model with highest probability			
$B_x\|x_m$	0.433	0.343	−0.266	−0.016
s	0.454	0.101	0.121	0.050
	Moderation term (if any)			
$B_{SP\times R\&D}$	12.933			−1.83
s	2.262			0.605
	Probability weighted estimate – Marginal effect at mean			
B_w	0.431	0.325	−0.252	−0.014
s	0.448	0.105	0.135	0.050

variance analyzed. When comparing firms, social performance appears to be positively related to financial performance, either directly or moderated through R&D. When comparing differences within a firm, social performance appears negatively, though less strongly, related to financial performance, either directly or moderated through R&D.

5 | DISCUSSION

We began this essay by arguing that it is often difficult to accumulate knowledge from quasi-replications, and our analysis reveals both a cause of this difficulty and a possible method for overcoming it. We show that quasi-replications are hard to interpret because their estimates are like snapshots of a broader empirical garden. We now use our broader perspective to provide a fuller picture of the SP–FP link and to consider the usefulness of model uncertainty analysis.

5.1 | Integrating the six studies

For the six studies, the empirical map reveals the critical importance of assumptions concerning between-firm and within-firm variance, and it highlights the danger of making comparisons

across this divide. Waddock and Graves's (1997) estimate was based on between-firm variance, but subsequent authors used both types, and this caused misunderstanding. For example, Zhao and Murrell (2016) report that W&G's finding "is not generalizable to a different sample that includes more firms over a longer time period" (Zhao & Murrell, 2016, p. 2386), but they fail to recognize that they use between-firm variance to evaluate their small sample and within-firm variance to evaluate their larger one. Our analysis suggests that had they used between-firm variance in the full sample, they would have confirmed W&G's result. Barnett and Salomon (2012) make a similar inferential misstep. They build up a series of increasingly complex models, culminating in one using within-firm variance, and based on the estimate from this last model, they infer that they only partially confirm W&G. In fact, they fully confirm it for the type of variance W&G used.

Our broader empirical map also helps us to place reported contingencies in context and thereby refine our understanding of their accuracy and importance. With our map in hand, we can see that M&S's evidence for R&D mediation is probably inaccurate or contingent on a particular empirical assumption (the use of a binary measure). We can also see that Hillman and Keim's (2001) claim for the value of scales emphasizing management action is probably overstated: in the broader context, the performance of their management scale appears comparable to those using the full KLD dataset. Finally, we can reasonably infer that Barnett and Salomon's (2012) U-shaped relationship between SP and FP is contingent on the use of between-firm variance and is better characterized as increasing returns.

The results of one study, Hull and Rothenberg (2008), are hard to reconcile with our broader empirical map. They find that R&D negatively moderates the relationship with social performance, but we find this effect only if we use within-firm variance, and doing so places us in a different part of the empirical garden. Perhaps we do not understand the assumptions H&R employed, or perhaps there is another empirical fork still to be mapped.

In total, with our map in hand, we can see that all of the studies are consistent with the inference that across firms social performance is positively correlated with financial performance, either directly or moderated by R&D. Over time within a firm, however, higher social performance seldom precedes higher financial return, either directly or moderated through R&D. Indeed, using some empirical assumptions, it is reasonable to infer a negative relationship between changes in social performance and subsequent financial performance.

5.2 | Future directions for research on the SP–FP link

Our evidence suggests explanations for the difficulty scholars have faced in trying to use meta-analytical techniques to synthesize pooled estimates of the SP–FP link. These studies must include estimates based on divergent empirical assumptions, such as the use of either between or within-firm variance, and thus the resulting meta-analytical estimates were influenced by the predominance of certain assumptions in the set of studies incorporated in the analysis. Thus, we provide evidence supporting Orlitzky's conjecture that meta-analysis of research on the SP–FP link is unlikely to deliver interpretable results unless great care is given to selecting studies that use similar empirical assumptions (Orlitzky, 2011, 2013).

Our analysis also suggests limited benefit from continued efforts to refine an ideal KLD-based measure of sustainability. Despite the efforts of these six authors to design better scales, the resulting measures were highly correlated and yielded similar coefficient estimates. This

observation is consistent with the finding by Berg et al. (2019) that elements of the KLD ratings are internally correlated.

Our analysis also reveals a puzzle that may help guide research on the SP–FP link: firms with high social performance also tend to have strong financial performance, but improvement in social performance seems, if anything, to reduce financial performance. Is some other factor, such as good management, influencing both? Are firm choices with respect to a sector or a competitive positioning part of the explanation? More research, perhaps incorporating model uncertainty analysis, is clearly needed.

Our greatest hope is that our presentation of model uncertainty analysis will cause researchers on the SP–FP link, or any empirical area, to be more thoughtful about their measurement and modeling assumptions, and more circumspect in interpreting the resulting estimates. As shown in Figure 2, the variance in the estimates across different models (i.e., epistemic uncertainty) usually exceeds the sampling variance of any given estimate (aleatoric uncertainty). In fact, the epistemic uncertainty shown in this graph considerably understates its full extent, because our analysis, based as it is on six similar studies, uses a relatively narrow set of assumptions with respect to the use of data, measures, and methods. Some of these data have been shown to contain errors and biases (Berg et al., 2019), and some of these methods have been superseded (Flammer, 2015). We hope that exposure to model uncertainty analysis will cause researchers to be more cautious about the boundaries of justified inference.

5.3 | Extensions and limitations of model uncertainty analysis

Our analysis also suggests directions for further research on the practical application of model uncertainty analysis. At present, the greatest challenge with the method involves the selection of the empirical assumptions to consider—what some scholars call "Occam's window" (Madigan & Raftery, 1994). This window should be broad enough to allow a good view but not so broad that it impedes analysis. At present, selecting a good window remains more of an art than a science. As Leamer noted a third of a century ago, the selection of a "credibly wide" window of assumptions will require judgment based on experience (Leamer, 1985). It is time we gathered that experience.

More research is also needed on how to present the results of a model uncertainty analysis. The graphs we use in this report allow the perception of broad patterns, but they do not allow an individual to index to estimates from a particular model. Simonsohn et al. (2020) suggest the use of codes on the *x*-axis to indicate what assumptions are in operation, but in our context, we found this approach unwieldy. One solution may be to provide an online database allowing readers to select assumptions and receive implied estimates.

Finally, like all research methods, model uncertainty analysis can be distorted or misapplied. One could, we suppose, manipulate the method to bring about a preferred pattern of results, but it is our subjective assessment that doing so would require considerable effort and ingenuity. We are more concerned that scholars may use the method in unwarranted ways; that scholars and readers may be tempted to find in the window of analysis answers to questions that have not even been posed. For example, despite our use of 3,680 models, our window of analysis provides no information on whether the relationship between social and financial performance is causal. None of the six studies we used in our analysis included assumptions that would allow identification of causality, so the answer to the question is simply not in view.

6 | CONCLUSION

Many scholars bemoan the difficulty of learning from individual research reports. Often, they propose that researchers should conduct more quasi-replications of existing research. Yet, even when replications are relatively common, their findings may prove difficult to form into a clear synthesis of shared understanding. As philosophers and statisticians have noted, empirical analysis often requires so many assumptions that it is like a walk through a garden with forking paths. Reports of where sincere scholars "came out" are hard to interpret unless more is known about the garden itself.

To demonstrate the problem and its possible solution, we use the case of six quasi-replications investigating the relationship between corporate social and financial performance. Although the studies use the same data sources and similar methods, they deliver conflicting estimates and interpretations. We show that quasi replications can be better understood once one is equipped with a map of the empirical garden that surrounds them. We demonstrate how such integrating analysis can be performed and presented.

The method we propose is rooted in a different theory of epistemology. In the traditional approach, scholars select and prespecify a few models to test, and from the estimates obtained, they advance objective and binary inferences. In the proposed approach, scholars define a set of reasonable models and then present, for the reader's consideration, estimates from a "window" of models using different assumptions.

Some readers main find our proposed method to be disconcerting. The epistemological approach we describe in this article requires scholars to be more precise and transparent about empirical uncertainties. It encourages the researcher to move, at least in part, from authority to guide. Yet, we found the experience to be liberating, not restricting. Rather than try to learn from light coming through a pinprick hole created by a few selected models, we could define a set of reasonable empirical assumptions and look through the window they created. We could evaluate and describe broad patterns we perceived, and yet still make transparent the uncertainty that remained.

It is our opinion that model uncertainty analysis should be added to the everyday empirical toolbox of management scholars. We can, as demonstrated in this article, use it to place in context a set of related studies, but we think we should use these new tools more often than that. We should use model uncertainty analysis *immediately* and *regularly*. Had modern computing power been available to the authors of the first study in our set, Waddock and Graves (1997), they could have begun the process of mapping this particular garden of forking paths. Imagine how that might have influenced the work of subsequent scholars. Imagine what progress we can make now if we make model uncertainty analysis a standard practice.

ACKNOWLEDGEMENTS

We are grateful to the many people that helped us with this article. We would especially like to thank David Rivers, Caroline Flammer, Brent Goldfarb, the two anonymous reviewers, and the Associate Editor Rajshree Agarwal. We gratefully acknowledge feedback from the participants at the Strategic Management Society 2020, ARCS Conference 2020, and GRONEN Reading Group 2020, as well as our colleagues at Rotterdam School of Management and Boston University. The usual disclaimer applies.

OPEN RESEARCH BADGES

This article has been awarded Open Data Badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. Data is available at Open Science Framework.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Luca Berchicci https://orcid.org/0000-0002-4285-9513
Andrew A. King https://orcid.org/0000-0003-4153-0326

REFERENCES

Achen, C. H. (2000). *Why lagged dependent variables can suppress the explanatory power of other independent variables.* Annual meeting of the political methodology section of the American political science association, UCLA (Vol. 20. No. 22).

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.

Albuquerque, R., Koskinen, Y., & Zhang, C. (2019). Corporate social responsibility and firm risk: Theory and empirical evidence. *Management Science*, *65*(10), 4451–4469.

Awaysheh, A., Heron, R. A., Perry, T., & Wilson, J. I. (2020). On the relation between corporate social responsibility and financial performance. *Strategic Management Journal*, *41*(6), 965–987.

Barnett, M. L., & Salomon, R. M. (2012). Does it pay to be really good? Addressing the shape of the relationship between social and financial performance. *Strategic Management Journal*, *33*(11), 1304–1320.

Bettis, R. A., Ethiraj, S., Gambardella, A., Helfat, C., & Mitchell, W. (2016). Creating repeatable cumulative knowledge in strategic management: A call for a broad and deep conversation among authors, referees, and editors. *Strategic Management Journal*, *37*(2), 257–261.

Bettis, R. A., Helfat, C. E., & Shaver, J. M. (2016). The necessity, logic, and forms of replication. *Strategic Management Journal*, *37*(11), 2193–2203.

Berg, F., Koelbel, J. F. & Rigobon, R. (2019). *Aggregate confusion: The divergence of ESG ratings*. Cambridge, Massachusetts: MIT Sloan School of Management.

Chamberlain, G., & Leamer, E. E. (1976). Matrix weighted averages and posterior bounds. *Journal of the Royal Statistical Society: Series B: Methodological*, *38*(1), 73–84.

Chevret, S., Ferguson, N. D., & Bellomo, R. (2018). Are systematic reviews and meta-analyses still useful research? *Intensive Care Medicine*, *44*, 515–517.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, *90* (432), 1313–1321.

De Long, J. B., & Lang, K. (1992). Are all economic hypotheses false? *Journal of Political Economy*, *100*(6), 1257–1272.

Durlauf, S. N., Navarro, S., & Rivers, D. A. (2016). Model uncertainty and the effect of shall-issue right-to-carry laws on crime. *European Economic Review*, *81*, 32–67.

Fernández, C., Ley, E., & Steel, M. F. J. (2001a). Benchmark priors for Bayesian model averaging. *Journal of Econometrics*, *100*(2), 381–427.

Fernández, C., Ley, E., & Steel, M. F. J. (2001b). Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics*, *16*(5), 563–576.

Flammer, C. (2015). Does corporate social responsibility lead to superior financial performance? A regression discontinuity approach. *Management Science*, *61*(11), 2549–2568.

Fricker, E. (2002). Trusting others in the sciences: A priori or empirical warrant? *Studies in History and Philosophy of Science Part A*, *33*(2), 373–383.

Gelman A., & Loken E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no 'fishing expedition' or 'p-hacking'. Unpublished draft.

Hacking, I. (2001). *An introduction to probability and inductive logic*. Cambridge, UK: Cambridge University Press.

Hacking, I. (2016). *The logic of statistical inference*. Cambridge, UK: Cambridge University Press.

Hillman, A. J., & Keim, G. D. (2001). Shareholder value, stakeholder management, and social issues: What's the bottom line? *Strategic Management Journal*, *22*(2), 125–139.

Horváthová, E. (2010). Does environmental performance affect financial performance? A meta-analysis. *Ecological Economics*, *70*(1), 52–59.

Hull, C. E., & Rothenberg, S. (2008). Firm performance: The interactions of corporate social performance with innovation and industry differentiation. *Strategic Management Journal*, *29*, 781–789.

Hume, D. (2000). *An enquiry concerning human understanding: A critical edition, year 1748* (Vol. 3). Oxford: Oxford University Press.

Keele, L., & Kelly, N. J. (2006). Dynamic models for dynamic theories: The ins and outs of lagged dependent variables. *Political Analysis*, *14*(2), 186–205.

King, A. A., Goldfarb, B., & Simcoe, T. (2021). Learning from testimony on quantitative research in management. *Academy of Management Review*, *46*(3), 465–488.

Leamer, E. E. (1982). Sets of posterior means with bounded variance priors. *Econometrica: Journal of the Econometric Society*, *50*, 725–736.

Leamer, E. E. (1985). Sensitivity analyses would help. *American Economic Review*, *75*(3), 308–313.

Leamer, E. E. (2010). Extreme bounds analysis. In S. N. Durlauf & L. E. Blume (Eds.), *Microeconometrics* (pp. 49–52). London: Palgrave Macmillan.

Levine, R., & Renelt, D. (1992). A sensitivity analysis of cross-country growth regressions. *The American Economic Review*, *82*, 942–963.

Longino, H. E. (1990). *Science as social knowledge: Values and objectivity in scientific inquiry*. Princeton, NJ: Princeton University Press.

Longino, H. E. (2019). The social dimensions of scientific knowledge. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Stanford, CA: Metaphysics Research Lab, Stanford University.

Madigan, D., & Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, *89*(428), 1535–1546.

Margolis, J. D., & Walsh, J. P. (2003). Misery loves companies: Rethinking social initiatives by business. *Administrative Science Quarterly*, *48*(2), 268–305.

McWilliams, A., & Siegel, D. (2000). Corporate social responsibility and financial performance: Correlation or misspecification? *Strategic Management Journal*, *21*(5), 603–609.

Møller, M. H., Ioannidis, J. P. A., & Darmon, M. (2018). Are systematic reviews and meta-analyses still useful research? We are not sure. *Intensive Care Medicine*, *44*(4), 518–520.

Orlitzky, M. (2011). Institutional logics in the study of organizations: The social construction of the relationship between corporate social and financial performance. *Business Ethics Quarterly*, *21*(3), 409–444.

Orlitzky, M. (2013). Corporate social responsibility, noise, and stock market volatility. *Academy of Management Perspectives*, *27*(3), 238–254.

Orlitzky, M., Schmidt, F. L., & Rynes, S. L. (2003). Corporate social and financial performance: A meta-analysis. *Organization Studies*, *24*(3), 403–441.

Sala-i-Martin, X. X. (1997). *I just ran four million regressions*. Cambridge, MA: National Bureau of Economic Research.

Shaver, J. M. (2019). Interpreting interactions in linear fixed-effect regression models: When fixed-effect estimates are no longer within-effects. *Strategy Science*, *4*(1), 25–40.

Simonsohn U., Simmons J. P., & Nelson L. D. (2020). Specification curve: Descriptive and inferential statistics on all reasonable specifications. Available at *SSRN* 2694998.

Waddock, S. A., & Graves, S. B. (1997). The corporate social performance-financial performance link. *Strategic Management Journal*, *18*(4), 303–319.

Wilholt, T. (2013). Epistemic trust in science. *British Journal for the Philosophy of Science*, *64*(2), 233–253.

Zhao, X., & Murrell, A. J. (2016). Revisiting the corporate social performance-financial performance link: A replication of Waddock and Graves. *Strategic Management Journal*, *37*(11), 2378–2388.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Berchicci, L., & King, A. A. (2022). Building knowledge by mapping model uncertainty in six studies of social and financial performance. *Strategic Management Journal*, *43*(7), 1319–1346. https://doi.org/10.1002/smj.3374